Do Letter Frequencies Change Over Time?

INTRODUCTION

- My question is, “Do letter frequencies change over time?”
- I tested whether, and how much, letter frequencies change over time by taking English texts from different centuries, running them through a code I wrote in Python, and analyzing how many times each letter appears in relation to the total number of letters.
- I analyzed texts from the years 1600 AD to 2000 AD, taking 10 pieces of text from each century.

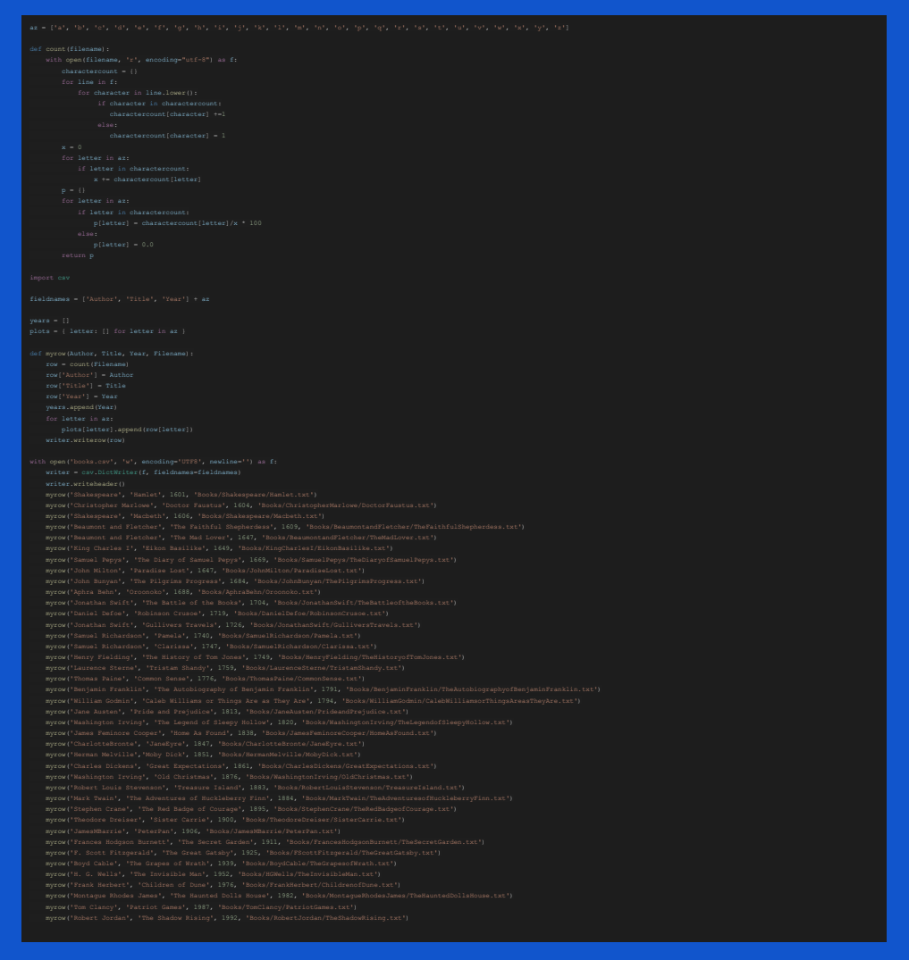
BACKGROUND RESEARCH

- Letter frequency is the ratio of the number of times a letter occurs in a text to the total number of letters in the text.
- For each letter, the correlation between the letter frequency and the year of the text can be measured using the Pearson coefficient and the p-value.
- The Pearson coefficient, or the correlation coefficient r, measures whether the correlation is strong or weak and positive or negative. It is positive when the slope of the "line" formed when the data is plotted on a graph is positive. It is negative when the slope of the line is negative.
- The stronger the correlation the more compact and neat of a line is formed. The correlation coefficient is always between -1 and 1. Weaker correlations are closer to 0 and stronger correlations are closer to -1 or 1.
- Using this correlation coefficient, the p-value can be calculated. The p-value is the probablitity that the results found, or more extreme results, would be obtained if there were no correlation.
- The p-value is always between 0 and 1. The closer to 0 it is, the less probable it is that there is actually no correlation and the more probable it is that there is a correlation. In statistics, a p-value less than 0.05 is usually considered "statistically significant."

HYPOTHESIS

- My hypothesis is that there are singnificant differences in letter frequencies in texts from different years.
- I think this because the English language evolves over time, so it is likely that the letter frequencies do too.

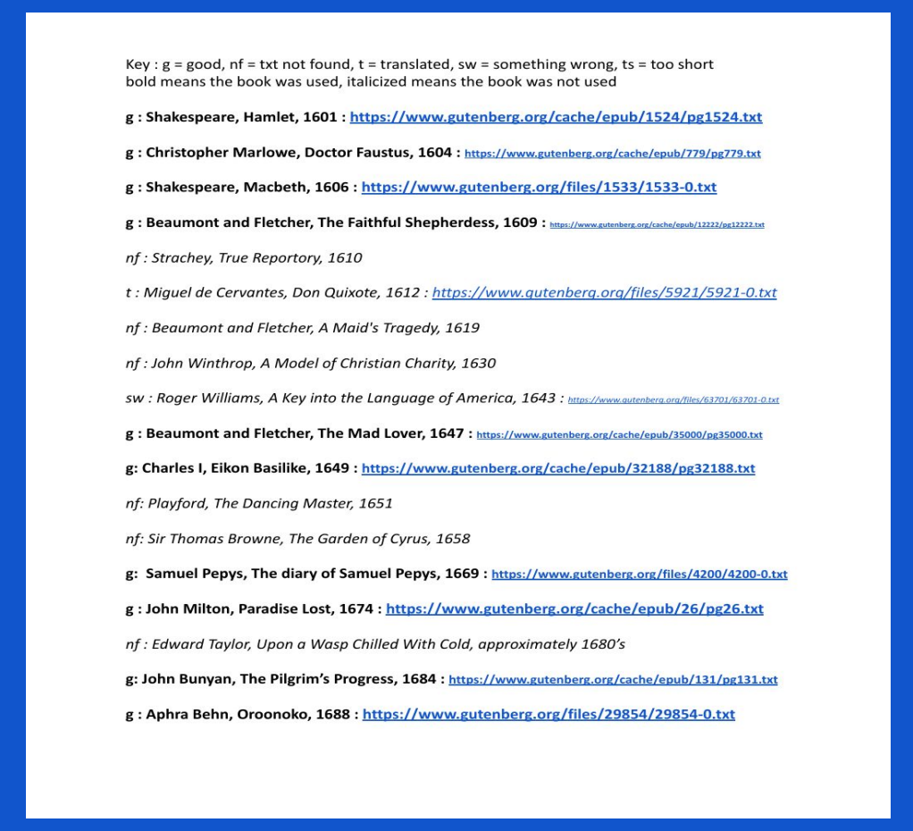
PROCEDURE



```
az = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z']

def count(filename):
    with open(filename, 'r', encoding="utf-8") as f:
        charactercount = {}
        for line in f:
            for character in line.lower():
                if character in charactercount:
                    charactercount[character] +=1
                else:
                    charactercount[character] = 1
        x = 0
        for letter in az:
            if letter in charactercount:
                x += charactercount[letter]
        p = {}
        for letter in az:
            if letter in charactercount:
                p[letter] = charactercount[letter]/x * 100
            else:
                p[letter] = 0.0
        return p

import csv

fieldnames = ['Author', 'Title', 'Year'] + az

years = []
plots = { letter: [] for letter in az }

def myrow(Author, Title, Year, Filename):
    row = count(Filename)
    row['Author'] = Author
    row['Title'] = Title
    row['Year'] = Year
    years.append(Year)
    for letter in az:
        plots[letter].append(row[letter])
    writer.writerow(row)

with open('books.csv', 'w', encoding='UTF8', newline='') as f:
    writer = csv.DictWriter(f, fieldnames=fieldnames)
    writer.writeheader()
    myrow('Shakespeare', 'Hamlet', 1601, 'Books/Shakespeare/Hamlet.txt')
    myrow('Christopher Marlowe', 'Doctor Faustus', 1604, 'Books/ChristopherMarlowe/DoctorFaustus.txt')
    myrow('Shakespeare', 'Macbeth', 1606, 'Books/Shakespeare/Macbeth.txt')
    myrow('Beaumont and Fletcher', 'The Faithful Shepherdess', 1609, 'Books/BeaumontandFletcher/TheFaithfulShepherdess.txt')
    myrow('Beaumont and Fletcher', 'The Mad Lover', 1647, 'Books/BeaumontandFletcher/TheMadLover.txt')
    myrow('King Charles I', 'Eikon Basilike', 1649, 'Books/KingCharlesI/EikonBasilike.txt')
    myrow('Samuel Pepys', 'The Diary of Samuel Pepys', 1669, 'Books/SamuelPepys/TheDiaryofSamuelPepys.txt')
    myrow('John Milton', 'Paradise Lost', 1647, 'Books/JohnMilton/ParadiseLost.txt')
    myrow('John Bunyan', 'The Pilgrims Progress', 1684, 'Books/JohnBunyan/ThePilgrimsProgress.txt')
    myrow('Aphra Behn', 'Oroonoko', 1688, 'Books/AphraBehn/Oroonoko.txt')
    myrow('Jonathan Swift', 'The Battle of the Books', 1704, 'Books/JonathanSwift/TheBattleoftheBooks.txt')
    myrow('Daniel Defoe', 'Robinson Crusoe', 1719, 'Books/DanielDefoe/RobinsonCrusoe.txt')
    myrow('Jonathan Swift', 'Gullivers Travels', 1726, 'Books/JonathanSwift/GulliversTravels.txt')
    myrow('Samuel Richardson', 'Pamela', 1740, 'Books/SamuelRichardson/Pamela.txt')
    myrow('Samuel Richardson', 'Clarissa', 1747, 'Books/SamuelRichardson/Clarissa.txt')
    myrow('Henry Fielding', 'The History of Tom Jones', 1749, 'Books/HenryFielding/TheHistoryofTomJones.txt')
    myrow('Laurence Sterne', 'Tristam Shandy', 1759, 'Books/LaurenceSterne/TristamShandy.txt')
    myrow('Thomas Paine', 'Common Sense', 1776, 'Books/ThomasPaine/CommonSense.txt')
    myrow('Benjamin Franklin', 'The Autobiography of Benjamin Franklin', 1791, 'Books/BenjaminFranklin/TheAutobiographyofBenjaminFranklin.txt')
    myrow('William Godmin', 'Caleb Williams or Things Are as They Are', 1794, 'Books/WilliamGodmin/CalebWilliamsorThingsAreasTheyAre.txt')
    myrow('Jane Austen', 'Pride and Prejudice', 1813, 'Books/JaneAusten/PrideandPrejudice.txt')
    myrow('Washington Irving', 'The Legend of Sleepy Hollow', 1820, 'Books/WashingtonIrving/TheLegendofSleepyHollow.txt')
    myrow('James Feminore Cooper', 'Home As Found', 1838, 'Books/JamesFeminoreCooper/HomeAsFound.txt')
    myrow('CharlotteBronte', 'JaneEyre', 1847, 'Books/CharlotteBronte/JaneEyre.txt')
    myrow('Herman Melville','Moby Dick', 1851, 'Books/HermanMelville/MobyDick.txt')
    myrow('Charles Dickens', 'Great Expectations', 1861, 'Books/CharlesDickens/GreatExpectations.txt')
    myrow('Washington Irving', 'Old Christmas', 1876, 'Books/WashingtonIrving/OldChristmas.txt')
    myrow('Robert Louis Stevenson', 'Treasure Island', 1883, 'Books/RobertLouisStevenson/TreasureIsland.txt')
    myrow('Mark Twain', 'The Adventures of Huckleberry Finn', 1884, 'Books/MarkTwain/TheAdventuresofHuckleberryFinn.txt')
    myrow('Stephen Crane', 'The Red Badge of Courage', 1895, 'Books/StephenCrane/TheRedBadgeofCourage.txt')
    myrow('Theodore Dreiser', 'Sister Carrie', 1900, 'Books/TheodoreDreiser/SisterCarrie.txt')
    myrow('JamesMBarrie', 'PeterPan', 1906, 'Books/JamesMBarrie/PeterPan.txt')
    myrow('Frances Hodgson Burnett', 'The Secret Garden', 1911, 'Books/FrancesHodgsonBurnett/TheSecretGarden.txt')
    myrow('F. Scott Fitzgerald', 'The Great Gatsby', 1925, 'Books/FScottFitzgerald/TheGreatGatsby.txt')
    myrow('Boyd Cable', 'The Grapes of Wrath', 1939, 'Books/BoydCable/TheGrapesofWrath.txt')
    myrow('H. G. Wells', 'The Invisible Man', 1952, 'Books/HGWells/TheInvisibleMan.txt')
    myrow('Frank Herbert', 'Children of Dune', 1976, 'Books/FrankHerbert/ChildrenofDune.txt')
    myrow('Montague Rhodes James', 'The Haunted Dolls House', 1982, 'Books/MontagueRhodesJames/TheHauntedDollsHouse.txt')
    myrow('Tom Clancy', 'Patriot Games', 1987, 'Books/TomClancy/PatriotGames.txt')
    myrow('Robert Jordan', 'The Shadow Rising', 1992, 'Books/RobertJordan/TheShadowRising.txt')
```

1. I wrote a code in Python that could calculate letter frequencies.

2. I made a list of books and found txt files of them. I downloaded the txt files and cleaned them up.



Key : g = good, nf = txt not found, t = translated, sw = something wrong, ts = too short
bold means the book was used, italicized means the book was not used

g : Shakespeare, Hamlet, 1601 : https://www.gutenberg.org/cache/epub/1524/pg1524.txt

g : Christopher Marlowe, Doctor Faustus, 1604 : https://www.gutenberg.org/cache/epub/779/pg779.txt

g : Shakespeare, Macbeth, 1606 : https://www.gutenberg.org/files/1533/1533-0.txt

g : Beaumont and Fletcher, The Faithful Shepherdess, 1609 : https://www.gutenberg.org/cache/epub/12222/pg12222.txt

nf : Strachey, True Reportory, 1610

t : Miguel de Cervantes, Don Quixote, 1612 : https://www.gutenberg.org/files/5921/5921-0.txt

nf : Beaumont and Fletcher, A Maid's Tragedy, 1619

nf : John Winthrop, A Model of Christian Charity, 1630

sw : Roger Williams, A Key into the Language of America, 1643 : https://www.gutenberg.org/files/63701/63701-0.txt

g : Beaumont and Fletcher, The Mad Lover, 1647 : https://www.gutenberg.org/cache/epub/35000/pg35000.txt

g: Charles I, Eikon Basilike, 1649 : https://www.gutenberg.org/cache/epub/32188/pg32188.txt

nf: Playford, The Dancing Master, 1651

nf: Sir Thomas Browne, The Garden of Cyrus, 1658

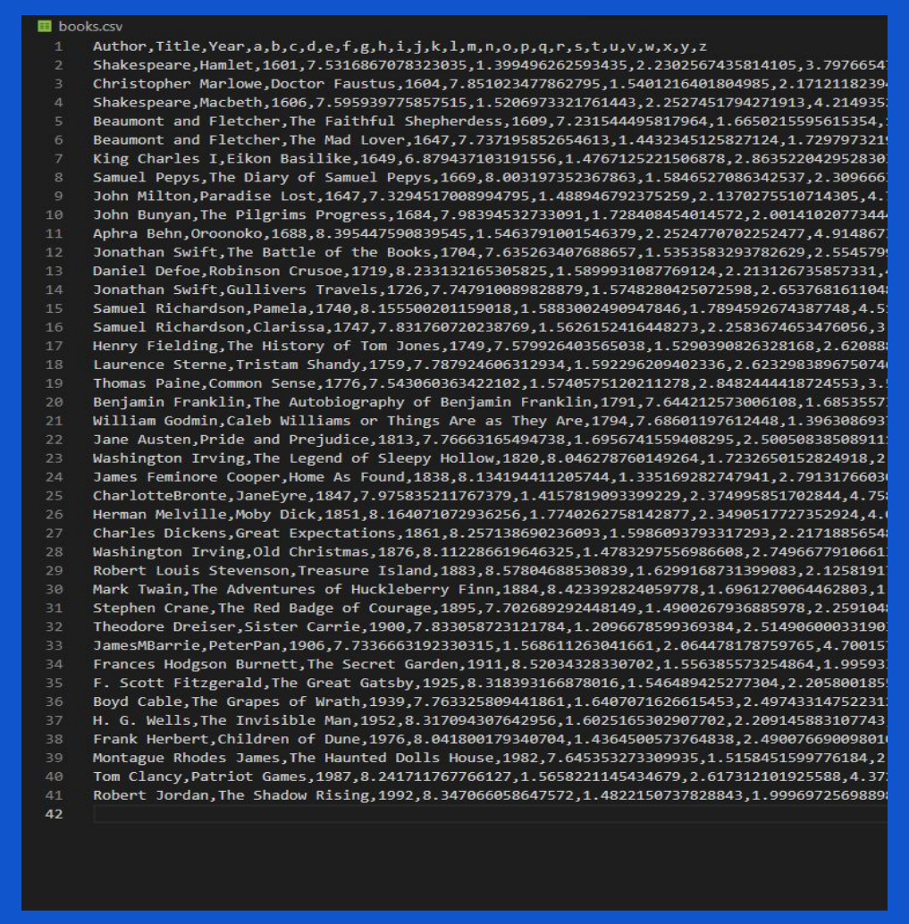
g: Samuel Pepys, The diary of Samuel Pepys, 1669 : https://www.gutenberg.org/files/4200/4200-0.txt

g : John Milton, Paradise Lost, 1674 : https://www.gutenberg.org/cache/epub/26/pg26.txt

nf : Edward Taylor, Upon a Wasp Chilled With Cold, approximately 1680’s

g: John Bunyan, The Pilgrim’s Progress, 1684 : https://www.gutenberg.org/cache/epub/131/pg131.txt

g : Aphra Behn, Oroonoko, 1688 : https://www.gutenberg.org/files/29854/29854-0.txt

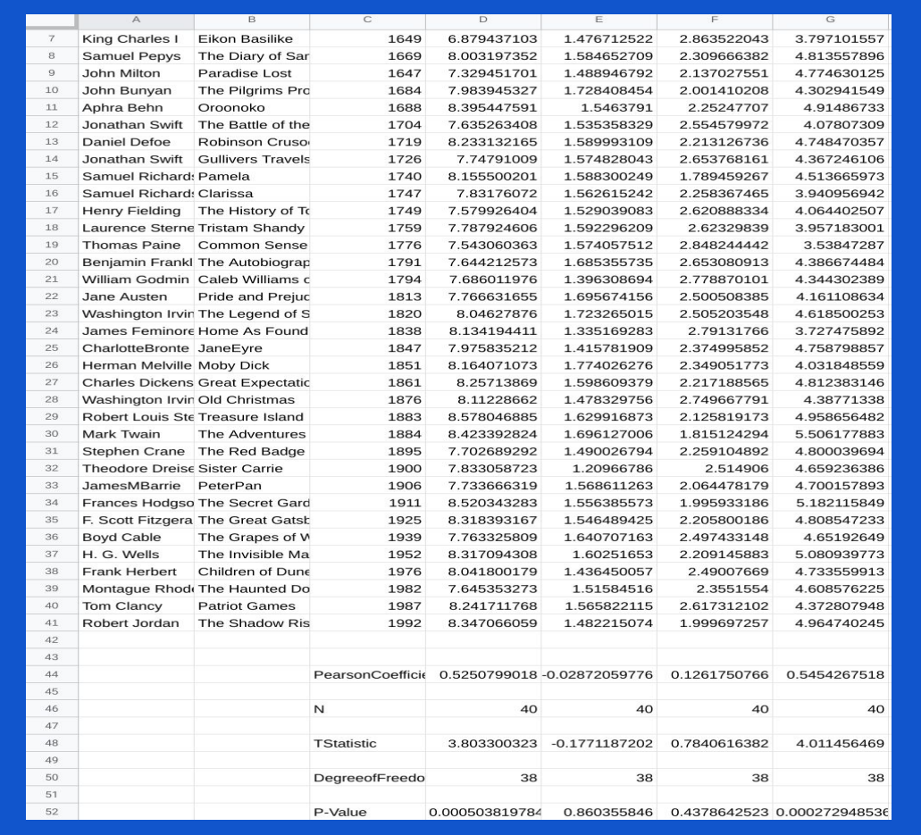


books.csv

```
Author,Title,Year,a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z
Shakespeare,Hamlet,1601,7.5316867078323035,1.399496262593435,2.2302567435814105,3.7976654
Christopher Marlowe,Doctor Faustus,1604,7.851023477862795,1.5401216401804985,2.17121182394
Shakespeare,Macbeth,1606,7.595939775857515,1.5206973321761443,2.2527451794271913,4.2149352
Beaumont and Fletcher,The Faithful Shepherdess,1609,7.231544495817964,1.6650215595615354,2
Beaumont and Fletcher,The Mad Lover,1647,7.737195852654613,1.4432345125827124,1.72979732l
King Charles I,Eikon Basilike,1649,6.879437103191556,1.4767125221506878,2.86352204295283O
Samuel Pepys,The Diary of Samuel Pepys,1669,8.003197352367863,1.5846527086342537,2.309666
John Milton,Paradise Lost,1647,7.3294517008994795,1.488946792375259,2.1370275510714305,4.7
John Bunyan,The Pilgrims Progress,1684,7.98394532733091,1.728408454014572,2.0014102077344
Aphra Behn,Oroonoko,1688,8.395447590839545,1.5463791001546379,2.2524770702252477,4.914867
Jonathan Swift,The Battle of the Books,1704,7.635263407688657,1.5353583293782629,2.554579
Daniel Defoe,Robinson Crusoe,1719,8.23313216530825,1.5899931087769124,2.21312673585733l,4
Jonathan Swift,Gullivers Travels,1726,7.747910089828879,1.5748280425072598,2.653768161104
Samuel Richardson,Pamela,1740,8.155500201159018,1.5883002490947846,1.7894592674387748,4.5
Samuel Richardson,Clarissa,1747,7.831760720238769,1.5626152416448273,2.2583674653476056,3
Henry Fielding,The History of Tom Jones,1749,7.579926403565038,1.5290390826328168,2.62088
Laurence Sterne,Tristam Shandy,1759,7.787924606312934,1.592296209402336,2.623298389675074
Thomas Paine,Common Sense,1776,7.543060363422102,1.5740575120211278,2.8482444418724553,3.5
Benjamin Franklin,The Autobiography of Benjamin Franklin,1791,7.644212573006108,1.6853557
William Godmin,Caleb Williams or Things Are as They Are,1794,7.68601197612448,1.396308693
Jane Austen,Pride and Prejudice,1813,7.76663165494738,1.6956741559408295,2.500508385089ll
Washington Irving,The Legend of Sleepy Hollow,1820,8.046278760149264,1.7232650152824918,2
James Feminore Cooper,Home As Found,1838,8.134194411205744,1.335169282747941,2.7913176603
CharlotteBronte,JaneEyre,1847,7.975835211767379,1.4157819093399229,2.374995851702844,4.75
Herman Melville,Moby Dick,1851,8.164071072936256,1.7740262758142877,2.3490517727352924,4.0
Charles Dickens,Great Expectations,1861,8.257138690236093,1.5986093793317293,2.2171885654
Washington Irving,Old Christmas,1876,8.112286619646325,1.4783297556986608,2.749667791066l
Robert Louis Stevenson,Treasure Island,1883,8.57804688530839,1.6299168731399083,2.1258191
Mark Twain,The Adventures of Huckleberry Finn,1884,8.423392824059778,1.6961270064462803,1
Stephen Crane,The Red Badge of Courage,1895,7.702689292448149,1.4900267936885978,2.259104
Theodore Dreiser,Sister Carrie,1900,7.833058723121784,1.2096678599369384,2.51490600033190
JamesMBarrie,PeterPan,1906,7.7336663192330315,1.568611263041661,2.064478178759765,4.70015
Frances Hodgson Burnett,The Secret Garden,1911,8.52034328330702,1.556385573254864,1.99593
F. Scott Fitzgerald,The Great Gatsby,1925,8.318393166878016,1.546489425277304,2.205800185
Boyd Cable,The Grapes of Wrath,1939,7.763325809441861,1.6407071626615453,2.49743314752231
H. G. Wells,The Invisible Man,1952,8.317094307642956,1.6025165302907702,2.209145883107743
Frank Herbert,Children of Dune,1976,8.041800179340704,1.4364500573764838,2.49007669009801
Montague Rhodes James,The Haunted Dolls House,1982,7.645353273309935,1.5158451599776184,2
Tom Clancy,Patriot Games,1987,8.241711767766127,1.5658221145434679,2.617312101925588,4.37
Robert Jordan,The Shadow Rising,1992,8.347066058647572,1.4822150737828843,1.99969725698
```

3. I ran my code to get the data. I uploaded the csv file with the data to a spreadsheet to analyze it.

4. I analyzed the data I collected from my trials by calculating the Pearson Coefficients and p-values.



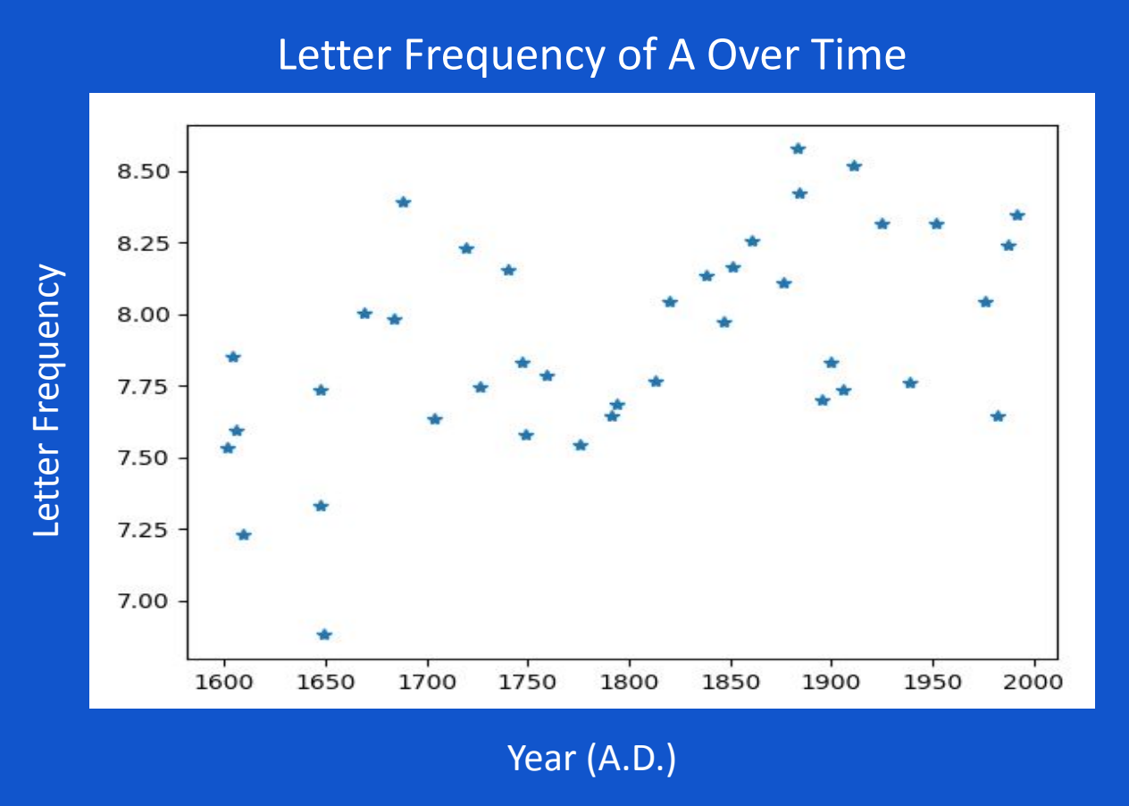
A	B	C	D	E	F	G
King Charles I	Eikon Basilike	1649	6.879437103	1.476712522	2.863522043	3.797101557
Samuel Pepys	The Diary of Sar	1669	8.003197352	1.584652709	2.309666382	4.813557896
John Milton	Paradise Lost	1647	7.329451701	1.488946792	2.137027551	4.774630125
John Bunyan	The Pilgrims Prc	1684	7.983945327	1.728408454	2.001410208	4.302941549
Aphra Behn	Oroonoko	1688	8.395447591	1.5463791	2.25247707	4.91486733
Jonathan Swift	The Battle of the	1704	7.635263408	1.535358329	2.554579972	4.07807309
Daniel Defoe	Robinson Cruso	1719	8.233132165	1.589993109	2.213126736	4.748470357
Jonathan Swift	Gullivers Travels	1726	7.74791009	1.574828043	2.653768161	4.367246106
Samuel Richard:	Pamela	1740	8.155500201	1.588300249	1.789459267	4.513665973
Samuel Richard:	Clarissa	1747	7.83176072	1.562615242	2.258367465	3.940956942
Henry Fielding	The History of Tc	1749	7.579926404	1.529039083	2.620888334	4.064402507
Laurence Sterne	Tristam Shandy	1759	7.787924606	1.592296209	2.62329839	3.957183001
Thomas Paine	Common Sense	1776	7.543060363	1.574057512	2.848244442	3.53847287
Benjamin Frankl	The Autobiograp	1791	7.644212573	1.685355735	2.653080913	4.386674484
William Godmin	Caleb Williams c	1794	7.686011976	1.396308694	2.778870101	4.344302389
Jane Austen	Pride and Prejuc	1813	7.766631655	1.695674156	2.500508385	4.161108634
Washington Irvin	The Legend of S	1820	8.04627876	1.723265015	2.505203548	4.618500253
James Feminore	Home As Found	1838	8.134194411	1.335169283	2.79131766	3.727475892
CharlotteBronte	JaneEyre	1847	7.975835212	1.415781909	2.374995852	4.758798857
Herman Melville	Moby Dick	1851	8.164071073	1.774026276	2.349051773	4.031848559
Charles Dickens	Great Expectatic	1861	8.25713869	1.598609379	2.217188565	4.812383146
Washington Irvin	Old Christmas	1876	8.11228662	1.478329756	2.749667791	4.38771338
Robert Louis Ste	Treasure Island	1883	8.578046885	1.629916873	2.125819173	4.958656482
Mark Twain	The Adventures	1884	8.423392824	1.696127006	1.815124294	5.506177883
Stephen Crane	The Red Badge	1895	7.702689292	1.490026794	2.259104892	4.800039694
Theodore Dreise	Sister Carrie	1900	7.833058723	1.20966786	2.514906	4.659236386
JamesMBarrie	PeterPan	1906	7.733666319	1.568611263	2.064478179	4.700157893
Frances Hodgso	The Secret Gard	1911	8.520343283	1.556385573	1.995933186	5.182115849
F. Scott Fitzgera	The Great Gatst	1925	8.318393167	1.546489425	2.205800186	4.808547233
Boyd Cable	The Grapes of W	1939	7.763325809	1.640707163	2.497433148	4.65192649
H. G. Wells	The Invisible Ma	1952	8.317094308	1.60251653	2.209145883	5.080939773
Frank Herbert	Children of Dune	1976	8.041800179	1.436450057	2.49007669	4.733559913
Montague Rhode	The Haunted Do	1982	7.645353273	1.51584516	2.3551554	4.608576225
Tom Clancy	Patriot Games	1987	8.241711768	1.565822115	2.617312102	4.372807948
Robert Jordan	The Shadow Ris	1992	8.347066059	1.482215074	1.999697257	4.964740245
		PearsonCoefficie	0.5250799018	-0.02872059776	0.1261750766	0.5454267518
		N	40	40	40	40
		TStatistic	3.803300323	-0.1771187202	0.7840616382	4.011456469
		DegreeofFreedo	38	38	38	38
		P-Value	0.000503819784	0.860355846	0.4378642523	0.00027294853€

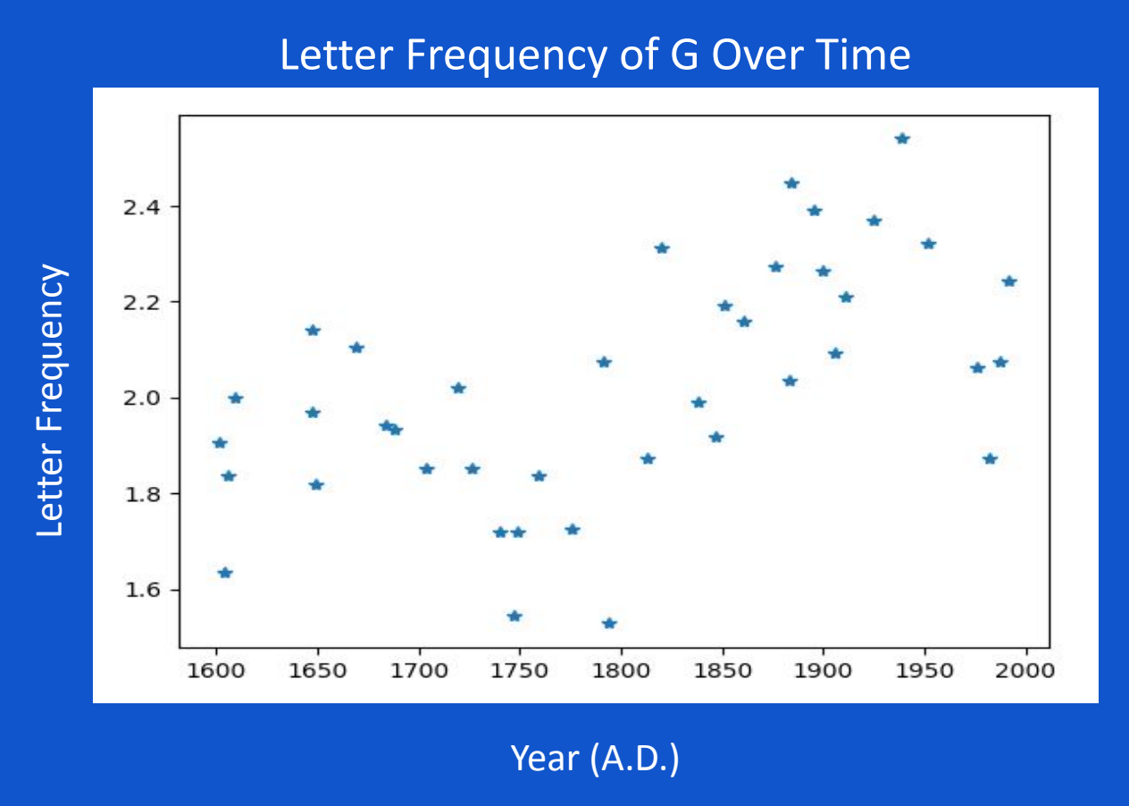
VARIABLES

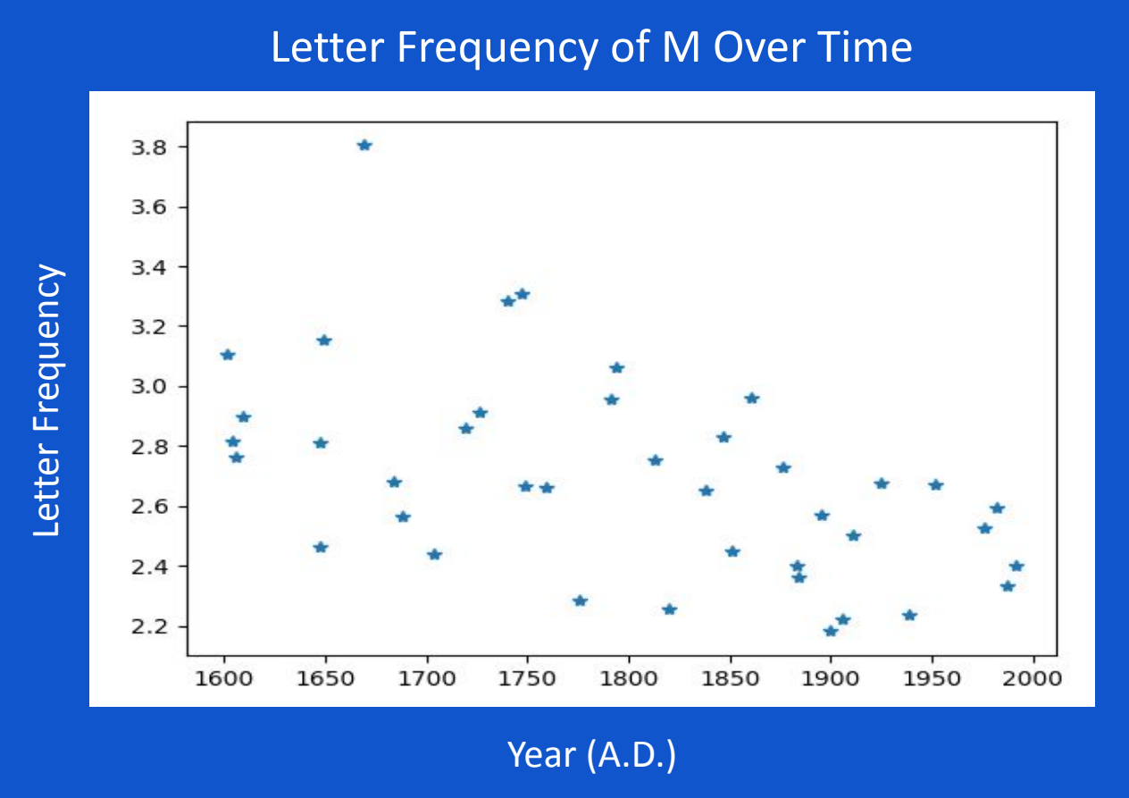
- Independent variable: year of text
- Dependent variable: letter frequency
- Constants: language and type of text

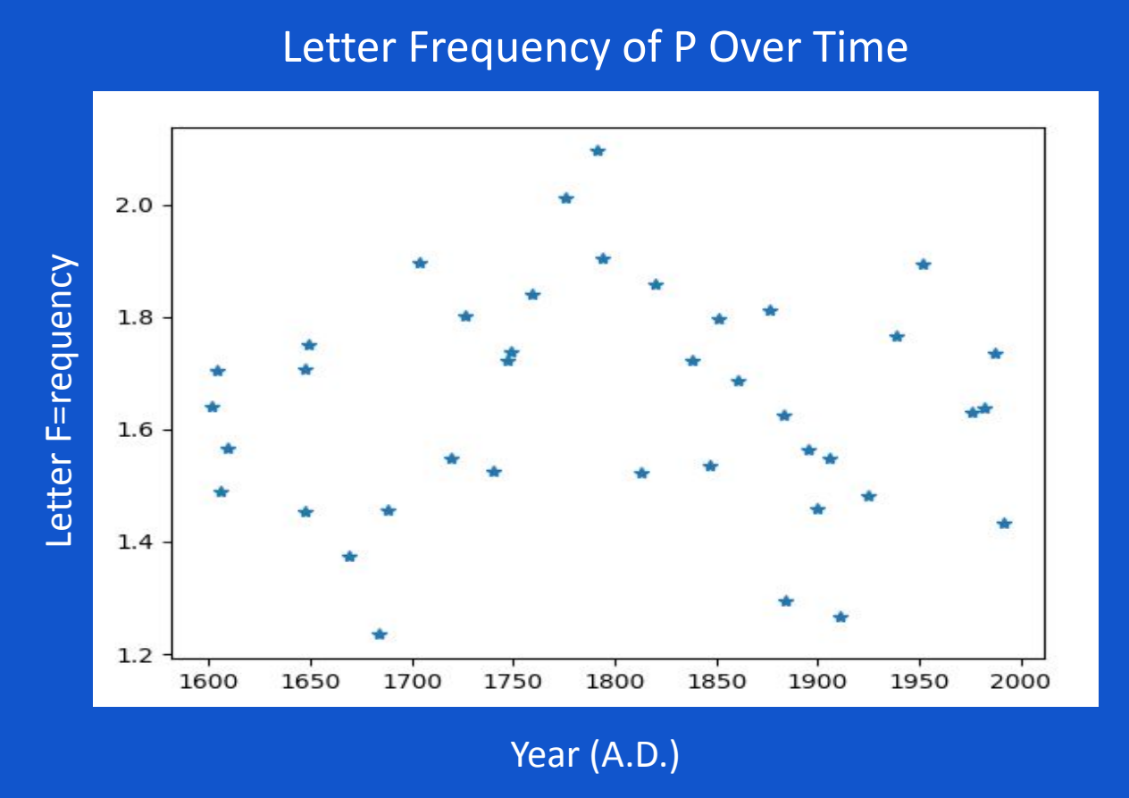
DATA

Letter	Correlation	P-value
a	0.525	0.0005
b	-0.029	0.86
c	0.126	0.438
d	0.545	0.0003
e	0.195	0.228
f	-0.267	0.95
g	0.546	0.0003
h	-0.048	0.678
i	0.051	0.754
j	0.242	0.133
k	0.375	0.017
l	-0.034	0.834
m	-0.519	0.0006
n	0.325	0.041
o	-0.449	0.004
p	0.011	0.946
q	0.065	0.688
r	0.39	0.013
s	-0.33	0.038
t	-0.043	0.79
u	-0.269	0.093
v	-0.524	0.0005
w	0.118	0.467
x	0.267	0.096
y	-0.263	0.101
z	0.206	0.202


Letter Frequency of A Over Time


Letter Frequency of G Over Time


Letter Frequency of M Over Time

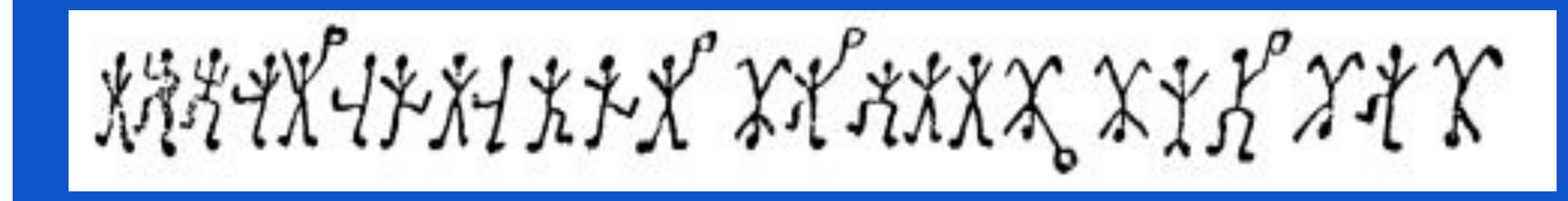

Letter Frequency of P Over Time

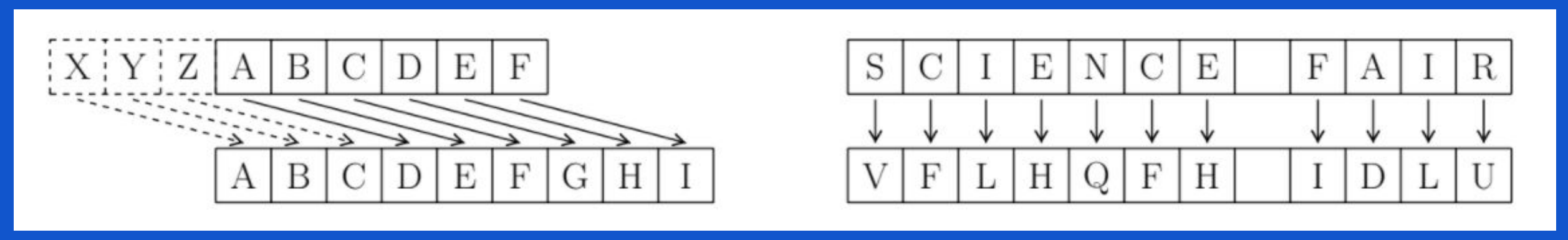
CONCLUSION

- My hypothesis was correct.
- My data shows that the frequency of ten letters (a, d, g, k, m, n, o, r, s, and v) change through time.
- Although not all of the letters have a significant correlation with time, enough do that the probability of none of the letters having a correlation is very small - much less than 5%, the accepted statistics probablity to assume that there is a correlation.
- Thus, both individually and all together, it is clear that there is a significant difference in letter frequencies in different centuries.

APPLICATION

- One application of letter frequency analysis is breaking codes. For example, it can be used for codes such as substitution cipher (where each letter is replaced by another letter or character).
- If there is a significant difference in the frequencies of any letters over time, then analyzing letter frequencies could also be used to identify the age of a text.
- Note that letter frequency analysis can be used with rather short texts as there are only twenty-six letters. On the other hand, word analysis can only be used with longer texts.





FURTHER INVESTIGATION

- One way to further this project is to analyze word frequencies. These will likely have greater differences over time than letter frequencies. The frequencies of different word lengths, sentence lengths, and punctuation can also be tested.
- Another idea is to analyze books that were translated from one language to another or books written in the same language but by authors with different primary languages.
- A third way is to analayze letter frequencies of different types of writting. Some examples are newspapers, scientific reports, and government documents.

WORKS CITED

- https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
- https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/
- https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/v/p-values-and-significance-tests/