ML EXERCISES

Group Members:

Utkarsh Garg
Eid: ug797

Aditya Sindhavad
Eid: aks5253

Ronak Goyal
Eid: rg49394

Agnitra Das
Eid: ad55985

Q1. Probability practice

Part A

Defining events –

- Event of clicking "yes" - Y
- Event of clicking "no" - N
- Event of truthful click - T
- Event of a random click – R

Probabilities given in problem statement-

- P(Y) = 0.65
- P(N) = 0.35
- P(R) = 0.3
- P(Y|R) = P(N|R) = 0.5

Using rule of total probability- P(Y) = P(Y,T) + P(Y,R)
P(Y) = P(Y|T).P(T) + P(Y|R).P(R)
0.65 = P(Y|T)x(1-0.3) + 0.5x0.3
P(Y|T) = 0.7143

Part B

- Event of having disease - D
- Event of not having disease - W
- Event of testing positive - P
- Event of testing negative – N

Probabilities given in problem statement-

- P(P|D) = 0.993
- P(N|W) = 0.9999
- P(D) = 0.000025

Using Baye's Theorem, P(D|P) = P(P|D).P(D)/P(P)

First, we need to find the value of P(P) using total probability.

P(P) = P(P|D).P(D) + P(P|W).P(W)

P(P) = (0.993)(0.000025) + (1-0.9999)(1-0.000025)
P(P) = 0.000125
Now, P(D|P) = (0.993x0.000025)/(0.000125) = 0.1989
P(D|P) = 0.1989

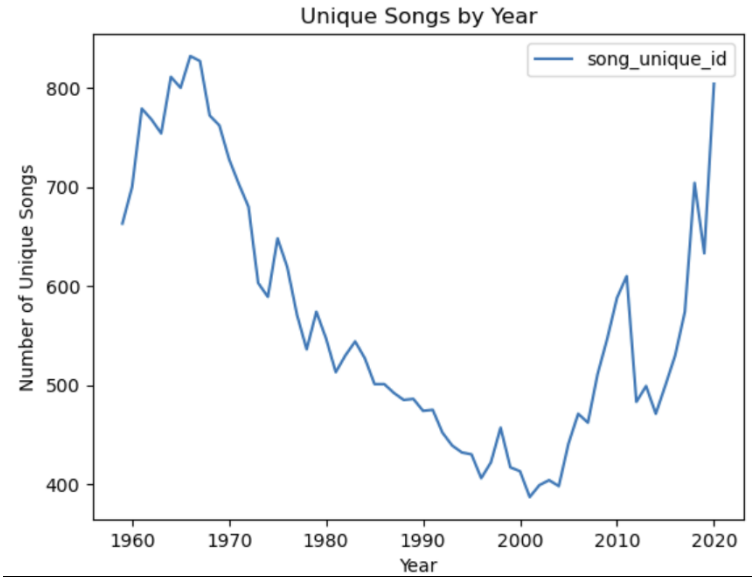
Q2. Wrangling the Billboard Top 100

Part A

		count
performer	song	
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
The Weeknd	Blinding Lights	76
Jason Mraz	I'm Yours	76
LeAnn Rimes	How Do I Live	69
OneRepublic	Counting Stars	68
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
Jewel	Foolish Games/You Were Meant For Me	65
Adele	Rolling In The Deep	65
Carrie Underwood	Before He Cheats	64

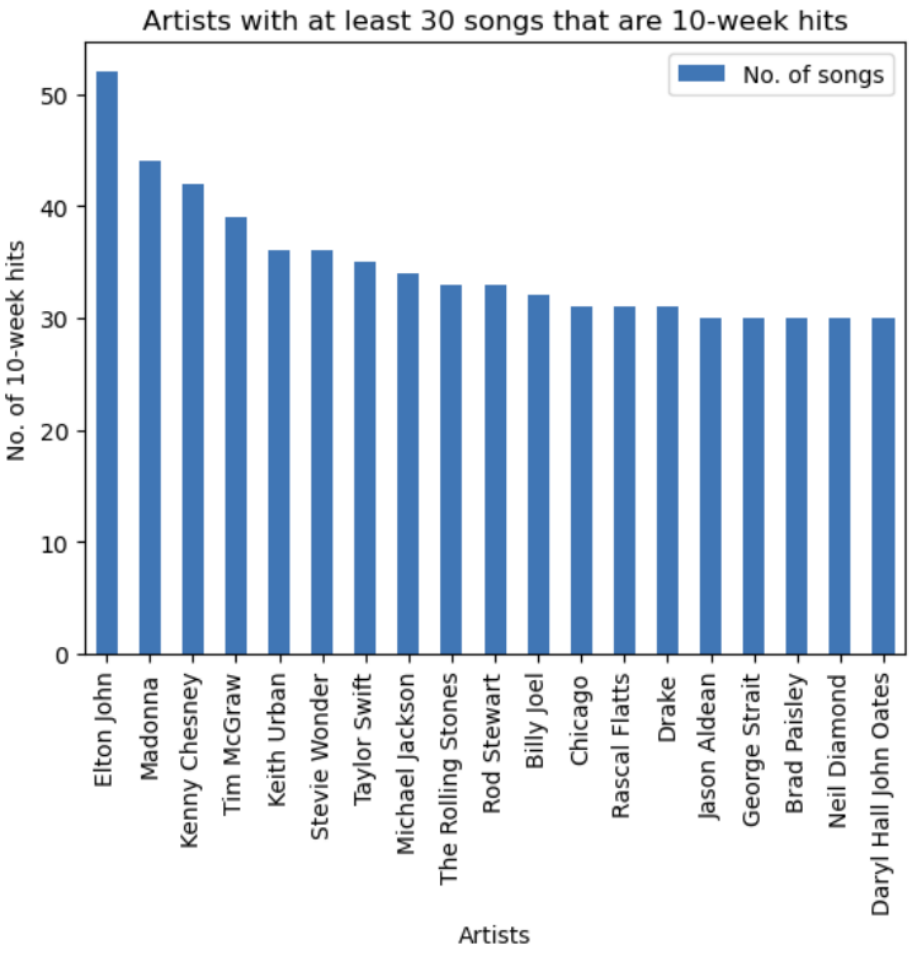
The table above gives us the top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100.

Part B



From the line chart above, we can observe that the "musical diversity" of the Billboard Top 100 was at its peak during the 1960s and then it kept decreasing for the next 40 years and hit its lowest in the 2000s. Post that, the diversity has again started to increase and the rate of increase is much higher than its decline before the 2000s.

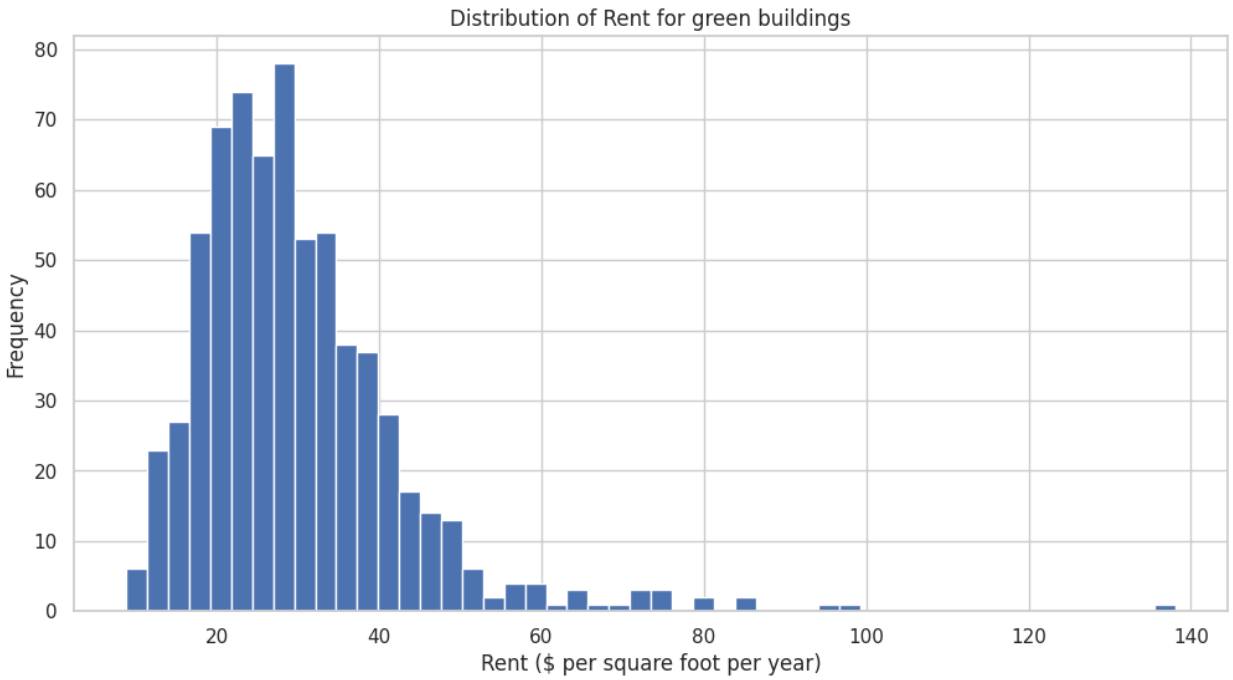
Part C

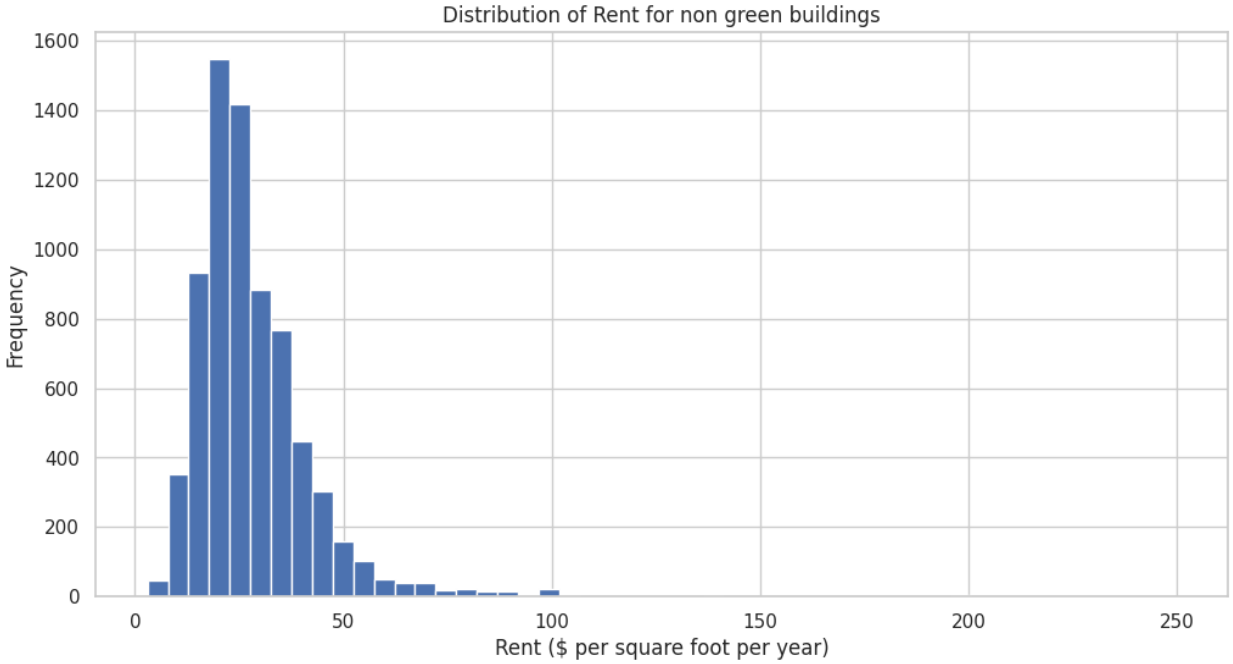

Artists with at least 30 songs that are 10-week hits
No. of songs
No. of 10-week hits
0
10
20
30
40
50
Elton John
Madonna
Kenny Chesney
Tim McGraw
Keith Urban
Stevie Wonder
Taylor Swift
Michael Jackson
The Rolling Stones
Rod Stewart
Billy Joel
Chicago
Rascal Flatts
Drake
Jason Aldean
George Strait
Brad Paisley
Neil Diamond
Daryl Hall John Oates
Artists

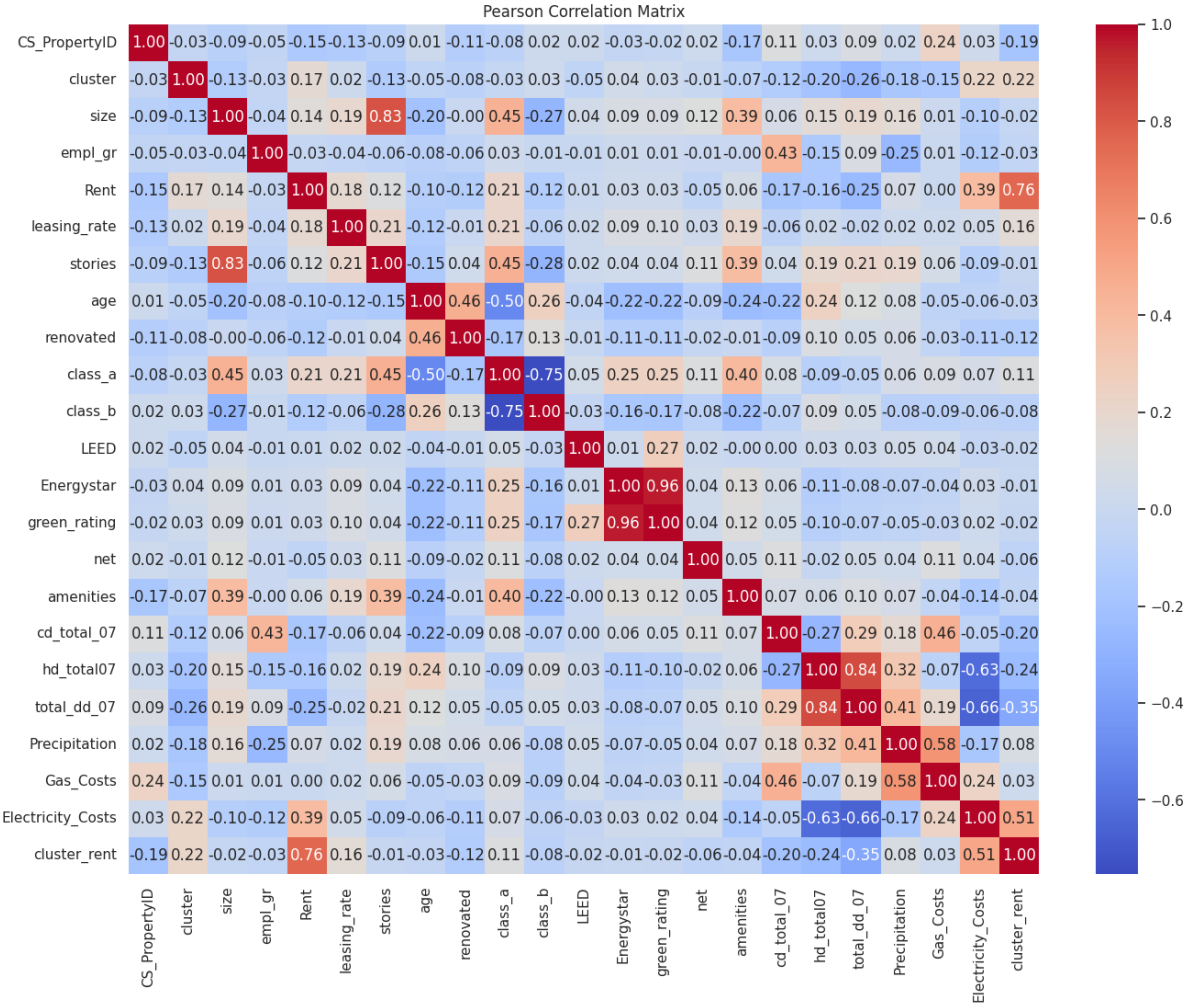
Q3. Visual storytelling part 1: green buildings

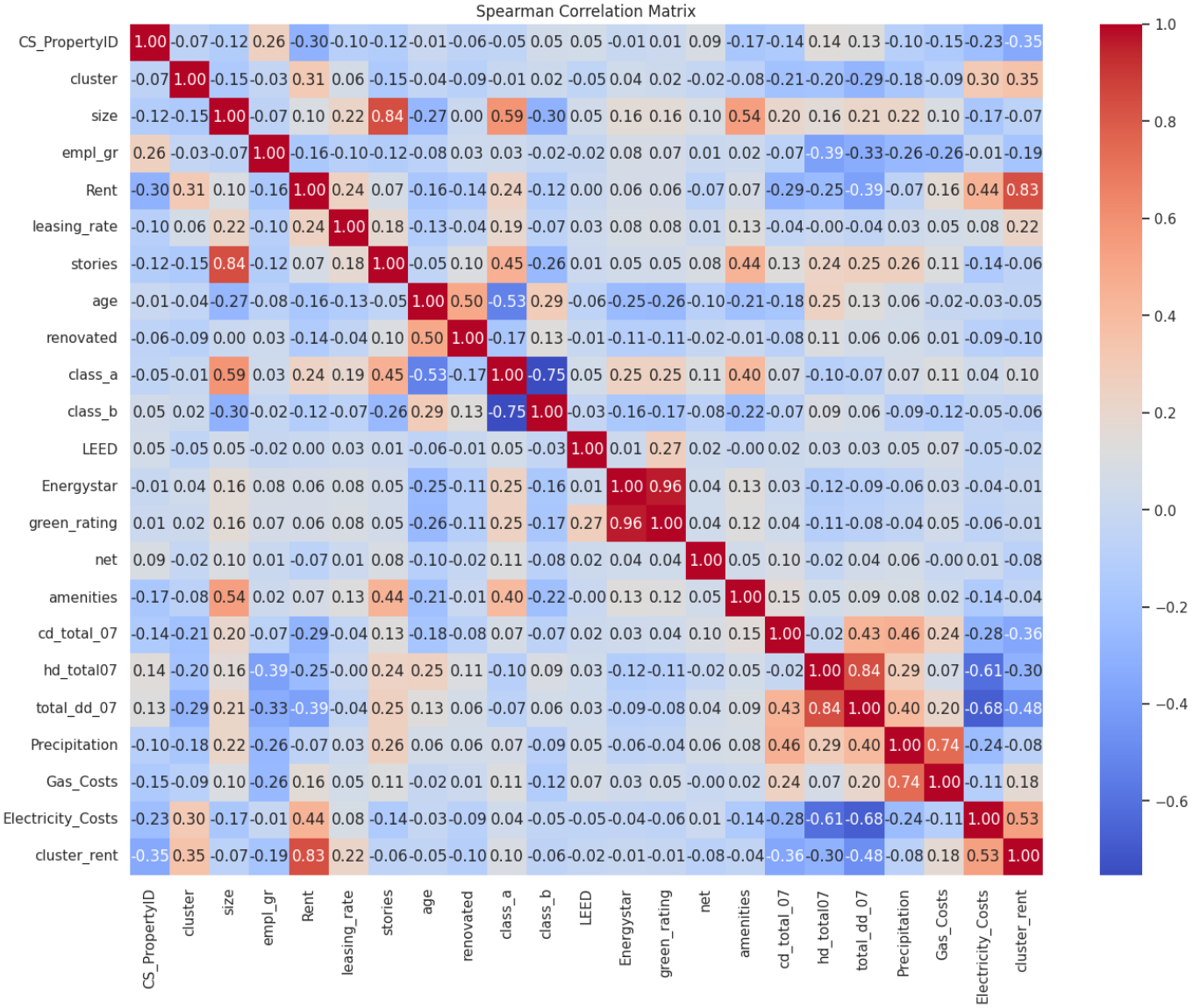
	Rent	LEED	Energystar	leasing_rate	size	age	cluster_rent
count	7894.000000	7894.000000	7894.000000	7894.000000	7.894000e+03	7894.000000	7894.000000
mean	28.418569	0.006841	0.080821	82.606371	2.346377e+05	47.243983	27.497285
std	15.075483	0.082430	0.272577	21.380315	2.975334e+05	32.194393	10.598952
min	2.980000	0.000000	0.000000	0.000000	1.624000e+03	0.000000	9.000000
25%	19.500000	0.000000	0.000000	77.850000	5.089125e+04	23.000000	20.000000
50%	25.160000	0.000000	0.000000	89.530000	1.288380e+05	34.000000	25.145000
75%	34.180000	0.000000	0.000000	96.440000	2.942120e+05	79.000000	34.000000
max	250.000000	1.000000	1.000000	100.000000	3.781045e+06	187.000000	71.440000

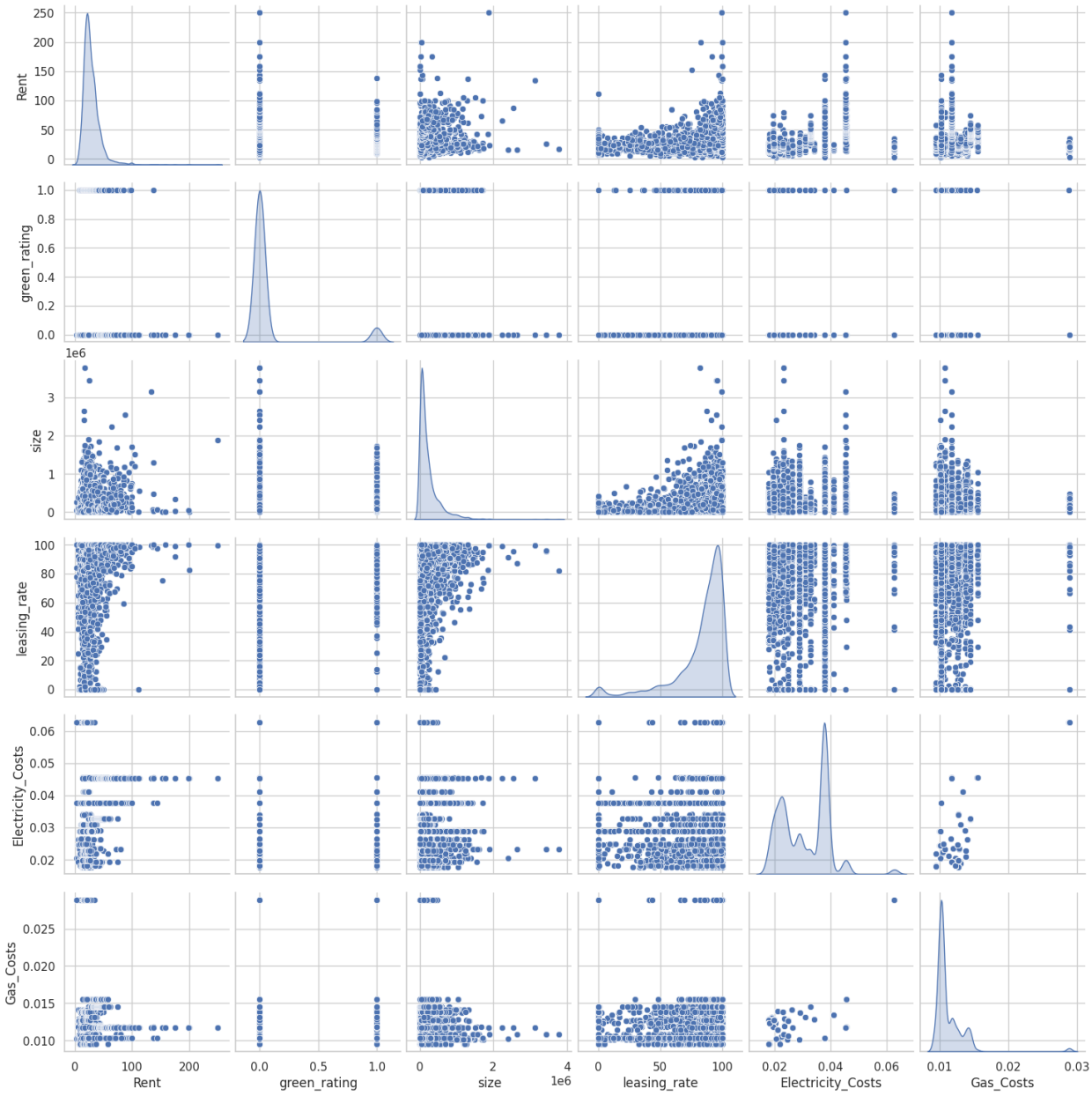
- Rent: The rent per square foot varies significantly, with a range of approximately $9 to $71.44
- Green Certification: Approximately 8.68% of the buildings are green-certified (either LEED or EnergyStar)
- Leasing Rate: The average leasing rate is around 82.6% while median is 89.53%
- Building Characteristics: Buildings vary widely in size, age, and other attributes

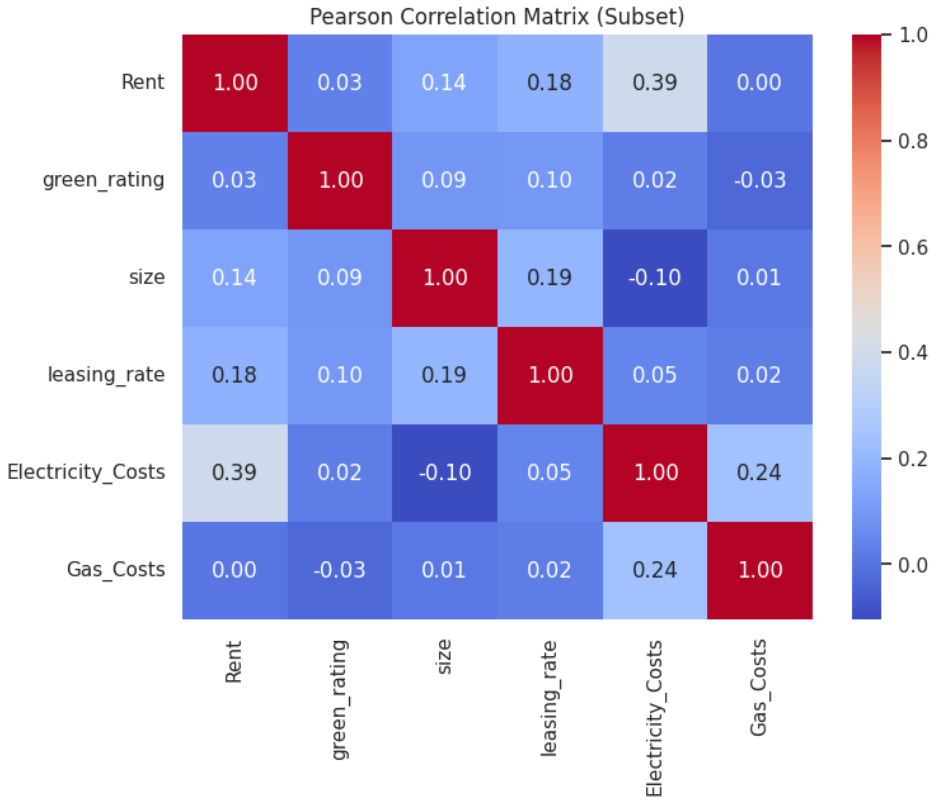


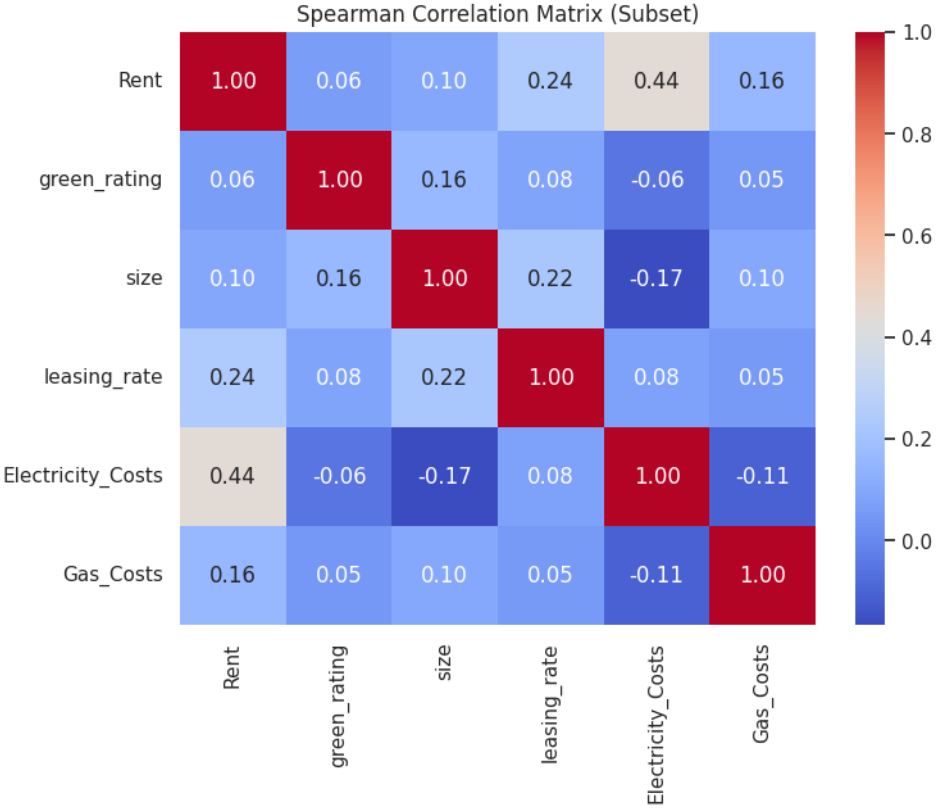

Distribution of Rent for non green buildings
Frequency
Rent ($ per square foot per year)
0
200
400
600
800
1000
1200
1400
1600
0
50
100
150
200
250

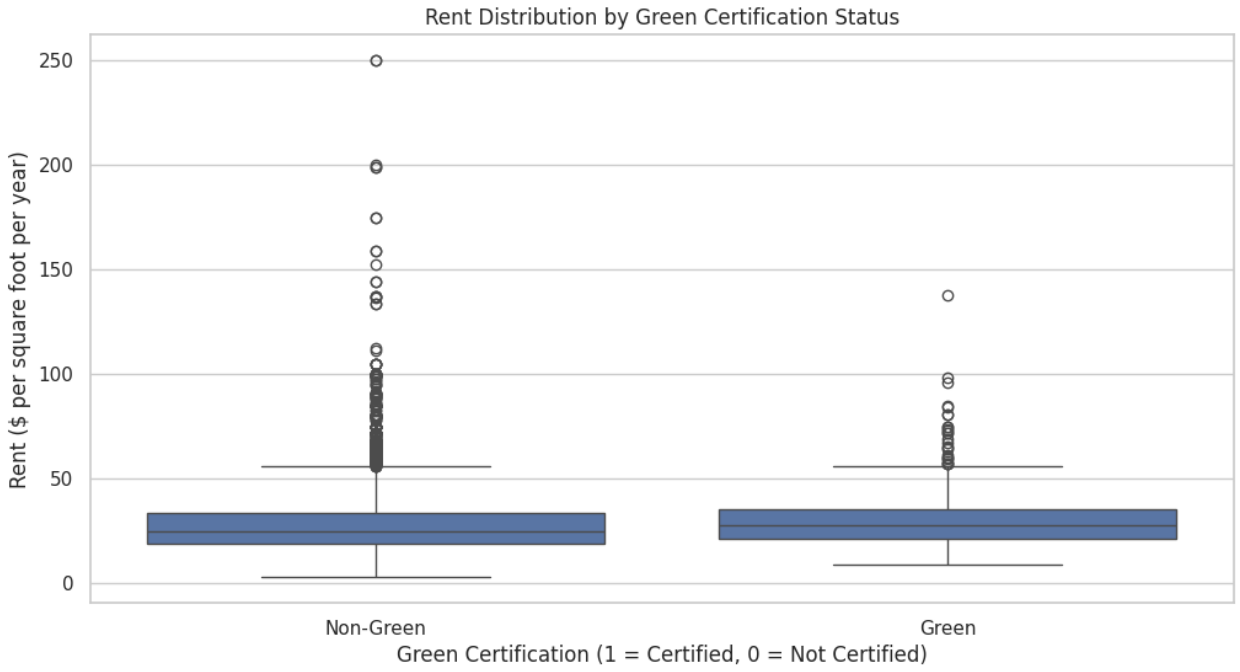

Pearson Correlation Matrix
CS_PropertyID 1.00 -0.03 -0.09 -0.05 -0.15 -0.13 -0.09 0.01 -0.11 -0.08 0.02 0.02 -0.03 -0.02 0.02 -0.17 0.11 0.03 0.09 0.02 0.24 0.03 -0.19
cluster -0.03 1.00 -0.13 -0.03 0.17 0.02 -0.13 -0.05 -0.08 -0.03 0.03 -0.05 0.04 0.03 -0.01 -0.07 -0.12 -0.20 -0.26 -0.18 -0.15 0.22 0.22
size -0.09 -0.13 1.00 -0.04 0.14 0.19 0.83 -0.20 -0.00 0.45 -0.27 0.04 0.09 0.09 0.12 0.39 0.06 0.15 0.19 0.16 0.01 -0.10 -0.02
empl_gr -0.05 -0.03 -0.04 1.00 -0.03 -0.04 -0.06 -0.08 -0.06 0.03 -0.01 -0.01 0.01 0.01 -0.01 -0.00 0.43 -0.15 0.09 -0.25 0.01 -0.12 -0.03
Rent -0.15 0.17 0.14 -0.03 1.00 0.18 0.12 -0.10 -0.12 0.21 -0.12 0.01 0.03 0.03 -0.05 0.06 -0.17 -0.16 -0.25 0.07 0.00 0.39 0.76
leasing_rate -0.13 0.02 0.19 -0.04 0.18 1.00 0.21 -0.12 -0.01 0.21 -0.06 0.02 0.09 0.10 0.03 0.19 -0.06 0.02 -0.02 0.02 0.02 0.05 0.16
stories -0.09 -0.13 0.83 -0.06 0.12 0.21 1.00 -0.15 0.04 0.45 -0.28 0.02 0.04 0.04 0.11 0.39 0.04 0.19 0.21 0.19 0.06 -0.09 -0.01
age 0.01 -0.05 -0.20 -0.08 -0.10 -0.12 -0.15 1.00 0.46 -0.50 0.26 -0.04 -0.22 -0.22 -0.09 -0.24 -0.22 0.24 0.12 0.08 -0.05 -0.06 -0.03
renovated -0.11 -0.08 -0.00 -0.06 -0.12 -0.01 0.04 0.46 1.00 -0.17 0.13 -0.01 -0.11 -0.11 -0.02 -0.01 -0.09 0.10 0.05 0.06 -0.03 -0.11 -0.12
class_a -0.08 -0.03 0.45 0.03 0.21 0.21 0.45 -0.50 -0.17 1.00 -0.75 0.05 0.25 0.25 0.11 0.40 0.08 -0.09 -0.05 0.06 0.09 0.07 0.11
class_b 0.02 0.03 -0.27 -0.01 -0.12 -0.06 -0.28 0.26 0.13 -0.75 1.00 -0.03 -0.16 -0.17 -0.08 -0.22 -0.07 0.09 0.05 -0.08 -0.09 -0.06 -0.08
LEED 0.02 -0.05 0.04 -0.01 0.01 0.02 0.02 -0.04 -0.01 0.05 -0.03 1.00 0.01 0.27 0.02 -0.00 0.00 0.03 0.03 0.05 0.04 -0.03 -0.02
Energystar -0.03 0.04 0.09 0.01 0.03 0.09 0.04 -0.22 -0.11 0.25 -0.16 0.01 1.00 0.96 0.04 0.13 0.06 -0.11 -0.08 -0.07 -0.04 0.03 -0.01
green_rating -0.02 0.03 0.09 0.01 0.03 0.10 0.04 -0.22 -0.11 0.25 -0.17 0.27 0.96 1.00 0.04 0.12 0.05 -0.10 -0.07 -0.05 -0.03 0.02 -0.02
net 0.02 -0.01 0.12 -0.01 -0.05 0.03 0.11 -0.09 -0.02 0.11 -0.08 0.02 0.04 0.04 1.00 0.05 0.11 -0.02 0.05 0.04 0.11 0.04 -0.06
amenities -0.17 -0.07 0.39 -0.00 0.06 0.19 0.39 -0.24 -0.01 0.40 -0.22 -0.00 0.13 0.12 0.05 1.00 0.07 0.06 0.10 0.07 -0.04 -0.14 -0.04
cd_total_07 0.11 -0.12 0.06 0.43 -0.17 -0.06 0.04 -0.22 -0.09 0.08 -0.07 0.00 0.06 0.05 0.11 0.07 1.00 -0.27 0.29 0.18 0.46 -0.05 -0.20
hd_total07 0.03 -0.20 0.15 -0.15 -0.16 0.02 0.19 0.24 0.10 -0.09 0.09 0.03 -0.11 -0.10 -0.02 0.06 -0.27 1.00 0.84 0.32 -0.07 -0.63 -0.24
total_dd_07 0.09 -0.26 0.19 0.09 -0.25 -0.02 0.21 0.12 0.05 -0.05 0.05 0.03 -0.08 -0.07 0.05 0.10 0.29 0.84 1.00 0.41 0.19 -0.66 -0.35
Precipitation 0.02 -0.18 0.16 -0.25 0.07 0.02 0.19 0.08 0.06 0.06 -0.08 0.05 -0.07 -0.05 0.04 0.07 0.18 0.32 0.41 1.00 0.58 -0.17 0.08
Gas_Costs 0.24 -0.15 0.01 0.01 0.00 0.02 0.06 -0.05 -0.03 0.09 -0.09 0.04 -0.04 -0.03 0.11 -0.04 0.46 -0.07 0.19 0.58 1.00 0.24 0.03
Electricity_Costs 0.03 0.22 -0.10 -0.12 0.39 0.05 -0.09 -0.06 -0.11 0.07 -0.06 -0.03 0.03 0.02 0.04 -0.14 -0.05 -0.63 -0.66 -0.17 0.24 1.00 0.51
cluster_rent -0.19 0.22 -0.02 -0.03 0.76 0.16 -0.01 -0.03 -0.12 0.11 -0.08 -0.02 -0.01 -0.02 -0.06 -0.04 -0.20 -0.24 -0.35 0.08 0.03 0.51 1.00
CS_PropertyID
cluster
size
empl_gr
Rent
leasing_rate
stories
age
renovated
class_a
class_b
LEED
Energystar
green_rating
net
amenities
cd_total_07
hd_total07
total_dd_07
Precipitation
Gas_Costs
Electricity_Costs
cluster_rent
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Spearman Correlation Matrix
CS_PropertyID 1.00 -0.07 -0.12 0.26 -0.30 -0.10 -0.12 -0.01 -0.06 -0.05 0.05 0.05 -0.01 0.01 0.09 -0.17 -0.14 0.14 0.13 -0.10 -0.15 -0.23 -0.35
cluster -0.07 1.00 -0.15 -0.03 0.31 0.06 -0.15 -0.04 -0.09 -0.01 0.02 -0.05 0.04 0.02 -0.02 -0.08 -0.21 -0.20 -0.29 -0.18 -0.09 0.30 0.35
size -0.12 -0.15 1.00 -0.07 0.10 0.22 0.84 -0.27 0.00 0.59 -0.30 0.05 0.16 0.16 0.10 0.54 0.20 0.16 0.21 0.22 0.10 -0.17 -0.07
empl_gr 0.26 -0.03 -0.07 1.00 -0.16 -0.10 -0.12 -0.08 0.03 0.03 -0.02 -0.02 0.08 0.07 0.01 0.02 -0.07 -0.39 -0.33 -0.26 -0.26 -0.01 -0.19
Rent -0.30 0.31 0.10 -0.16 1.00 0.24 0.07 -0.16 -0.14 0.24 -0.12 0.00 0.06 0.06 -0.07 0.07 -0.29 -0.25 -0.39 -0.07 0.16 0.44 0.83
leasing_rate -0.10 0.06 0.22 -0.10 0.24 1.00 0.18 -0.13 -0.04 0.19 -0.07 0.03 0.08 0.08 0.01 0.13 -0.04 -0.00 -0.04 0.03 0.05 0.08 0.22
stories -0.12 -0.15 0.84 -0.12 0.07 0.18 1.00 -0.05 0.10 0.45 -0.26 0.01 0.05 0.05 0.08 0.44 0.13 0.24 0.25 0.26 0.11 -0.14 -0.06
age -0.01 -0.04 -0.27 -0.08 -0.16 -0.13 -0.05 1.00 0.50 -0.53 0.29 -0.06 -0.25 -0.26 -0.10 -0.21 -0.18 0.25 0.13 0.06 -0.02 -0.03 -0.05
renovated -0.06 -0.09 0.00 0.03 -0.14 -0.04 0.10 0.50 1.00 -0.17 0.13 -0.01 -0.11 -0.11 -0.02 -0.01 -0.08 0.11 0.06 0.06 0.01 -0.09 -0.10
class_a -0.05 -0.01 0.59 0.03 0.24 0.19 0.45 -0.53 -0.17 1.00 -0.75 0.05 0.25 0.25 0.11 0.40 0.07 -0.10 -0.07 0.07 0.11 0.04 0.10
class_b 0.05 0.02 -0.30 -0.02 -0.12 -0.07 -0.26 0.29 0.13 -0.75 1.00 -0.03 -0.16 -0.17 -0.08 -0.22 -0.07 0.09 0.06 -0.09 -0.12 -0.05 -0.06
LEED 0.05 -0.05 0.05 -0.02 0.00 0.03 0.01 -0.06 -0.01 0.05 -0.03 1.00 0.01 0.27 0.02 -0.00 0.02 0.03 0.03 0.05 0.07 -0.05 -0.02
Energystar -0.01 0.04 0.16 0.08 0.06 0.08 0.05 -0.25 -0.11 0.25 -0.16 0.01 1.00 0.96 0.04 0.13 0.03 -0.12 -0.09 -0.06 0.03 -0.04 -0.01
green_rating 0.01 0.02 0.16 0.07 0.06 0.08 0.05 -0.26 -0.11 0.25 -0.17 0.27 0.96 1.00 0.04 0.12 0.04 -0.11 -0.08 -0.04 0.05 -0.06 -0.01
net 0.09 -0.02 0.10 0.01 -0.07 0.01 0.08 -0.10 -0.02 0.11 -0.08 0.02 0.04 0.04 1.00 0.05 0.10 -0.02 0.04 0.06 -0.00 0.01 -0.08
amenities -0.17 -0.08 0.54 0.02 0.07 0.13 0.44 -0.21 -0.01 0.40 -0.22 -0.00 0.13 0.12 0.05 1.00 0.15 0.05 0.09 0.08 0.02 -0.14 -0.04
cd_total_07 -0.14 -0.21 0.20 -0.07 -0.29 -0.04 0.13 -0.18 -0.08 0.07 -0.07 0.02 0.03 0.04 0.10 0.15 1.00 -0.02 0.43 0.46 0.24 -0.28 -0.36
hd_total07 0.14 -0.20 0.16 -0.39 -0.25 -0.00 0.24 0.25 0.11 -0.10 0.09 0.03 -0.12 -0.11 -0.02 0.05 -0.02 1.00 0.84 0.29 0.07 -0.61 -0.30
total_dd_07 0.13 -0.29 0.21 -0.33 -0.39 -0.04 0.25 0.13 0.06 -0.07 0.06 0.03 -0.09 -0.08 0.04 0.09 0.43 0.84 1.00 0.40 0.20 -0.68 -0.48
Precipitation -0.10 -0.18 0.22 -0.26 -0.07 0.03 0.26 0.06 0.06 0.07 -0.09 0.05 -0.06 -0.04 0.06 0.08 0.46 0.29 0.40 1.00 0.74 -0.24 -0.08
Gas_Costs -0.15 -0.09 0.10 -0.26 0.16 0.05 0.11 -0.02 0.01 0.11 -0.12 0.07 0.03 0.05 -0.00 0.02 0.24 0.07 0.20 0.74 1.00 -0.11 0.18
Electricity_Costs -0.23 0.30 -0.17 -0.01 0.44 0.08 -0.14 -0.03 -0.09 0.04 -0.05 -0.05 -0.04 -0.06 0.01 -0.14 -0.28 -0.61 -0.68 -0.24 -0.11 1.00 0.53
cluster_rent -0.35 0.35 -0.07 -0.19 0.83 0.22 -0.06 -0.05 -0.10 0.10 -0.06 -0.02 -0.01 -0.01 -0.08 -0.04 -0.36 -0.30 -0.48 -0.08 0.18 0.53 1.00
CS_PropertyID
cluster
size
empl_gr
Rent
leasing_rate
stories
age
renovated
class_a
class_b
LEED
Energystar
green_rating
net
amenities
cd_total_07
hd_total07
total_dd_07
Precipitation
Gas_Costs
Electricity_Costs
cluster_rent
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

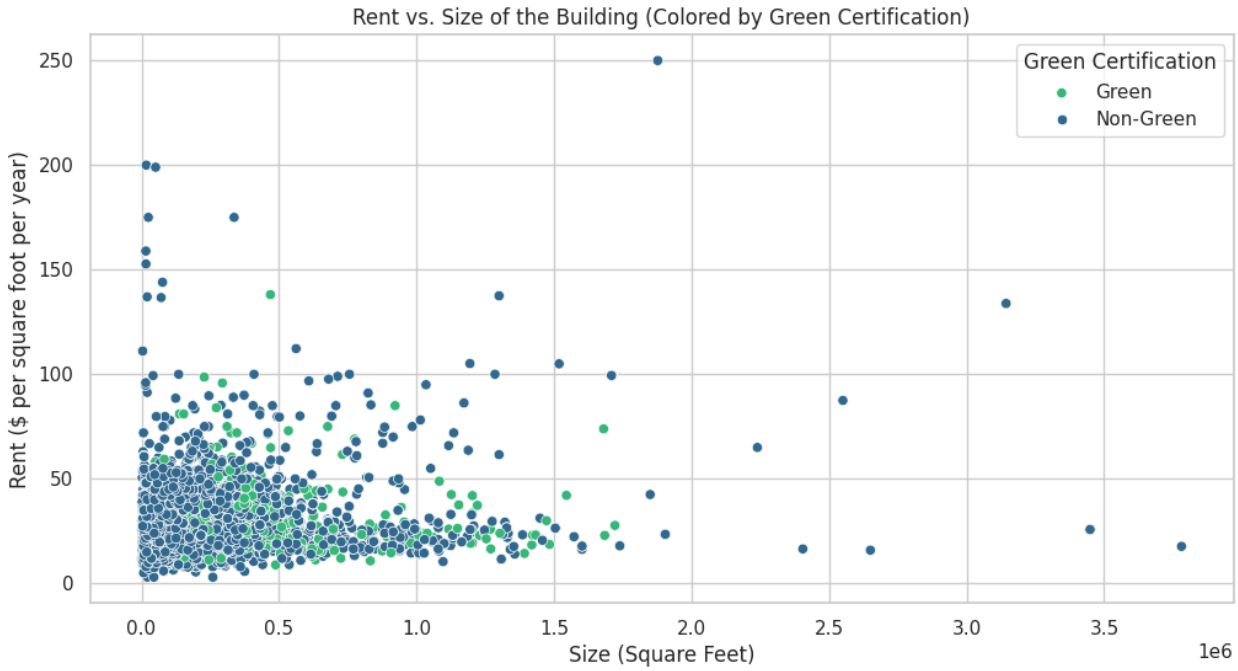

Rent
green_rating
size
1e6
leasing_rate
Electricity_Costs
Gas_Costs
Rent
green_rating
size
1e6
leasing_rate
Electricity_Costs
Gas_Costs

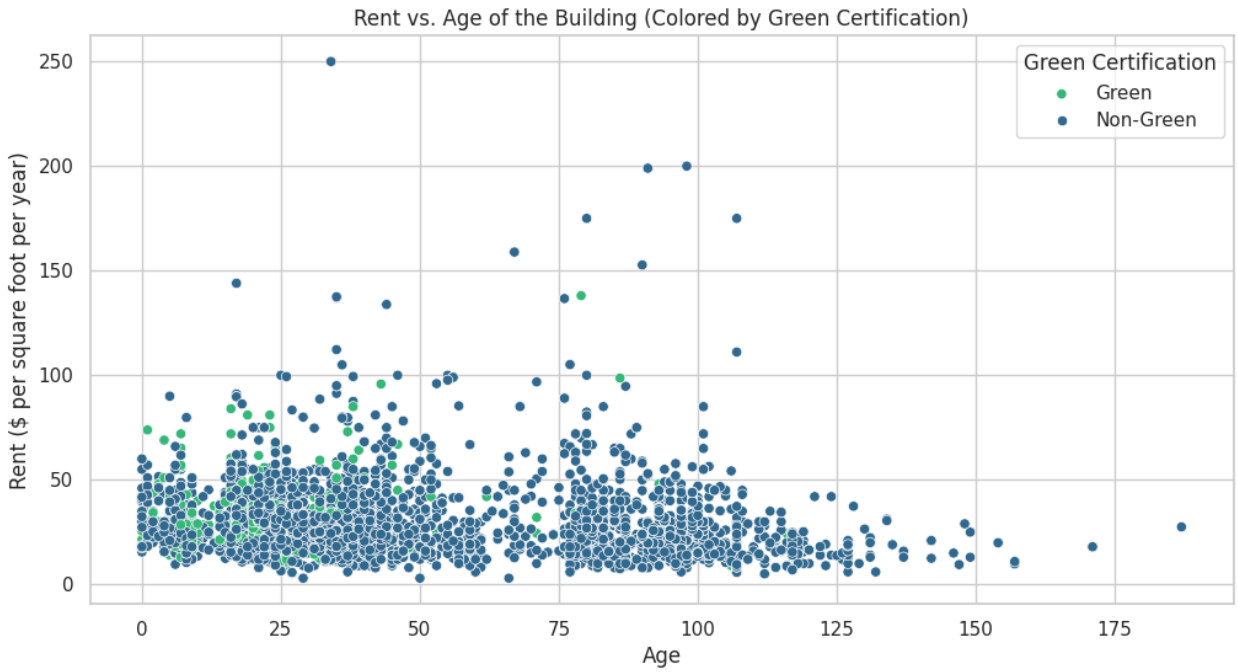

Pearson Correlation Matrix (Subset)
Rent
green_rating
size
leasing_rate
Electricity_Costs
Gas_Costs
1.00 0.03 0.14 0.18 0.39 0.00
0.03 1.00 0.09 0.10 0.02 -0.03
0.14 0.09 1.00 0.19 -0.10 0.01
0.18 0.10 0.19 1.00 0.05 0.02
0.39 0.02 -0.10 0.05 1.00 0.24
0.00 -0.03 0.01 0.02 0.24 1.00
Rent
green_rating
size
leasing_rate
Electricity_Costs
Gas_Costs
1.0
0.8
0.6
0.4
0.2
0.0


Spearman Correlation Matrix (Subset)
Rent
green_rating
size
leasing_rate
Electricity_Costs
Gas_Costs
1.00 0.06 0.10 0.24 0.44 0.16
0.06 1.00 0.16 0.08 -0.06 0.05
0.10 0.16 1.00 0.22 -0.17 0.10
0.24 0.08 0.22 1.00 0.08 0.05
0.44 -0.06 -0.17 0.08 1.00 -0.11
0.16 0.05 0.10 0.05 -0.11 1.00
Rent
green_rating
size
leasing_rate
Electricity_Costs
Gas_Costs
1.0
0.8
0.6
0.4
0.2
0.0



Green-certified buildings have higher median rent per square foot per year as compared to Non-Green Buildings





KEY FINDINGS:

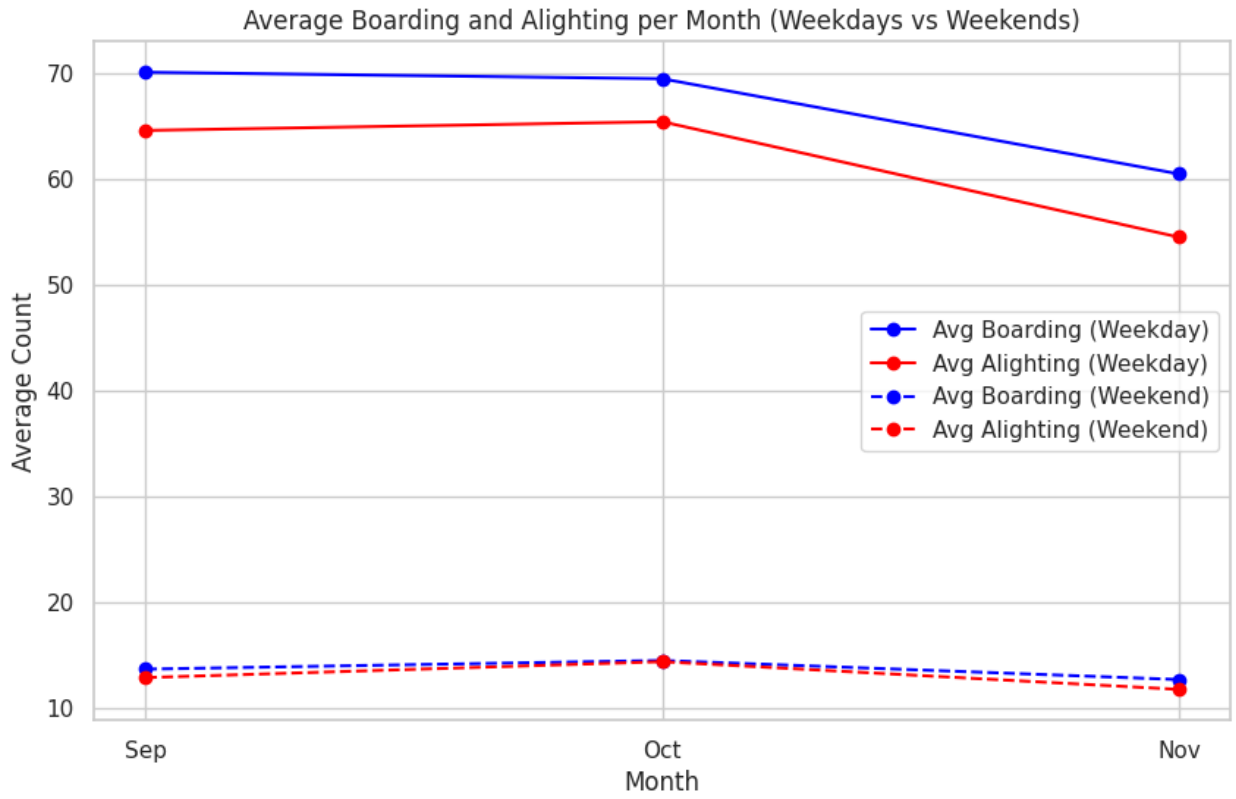
1. For a green building make sure it has the EnergyStar certification rather than any other certification
2. There is a minute trend between green certification and rent, suggesting that green buildings might command higher rents. However, the scatterplot indicates that the correlation could be stronger, and there are many non-green buildings with high rents as well.

3. This suggests that while green certification could contribute to higher rents, other factors are also at play, and green certification alone may not guarantee higher revenue, contradicting what was suggested by the guru in the earlier analysis
4. The plot shows a positive relationship between building size and rent, especially in larger buildings, which tend to have higher rents.
5. Buildings with higher occupancy rates tend to charge higher rents. Green buildings, being more attractive to sustainability-conscious tenants, may achieve higher occupancy, leading to increased and consistent revenue.
6. Buildings with lower electricity costs, often due to energy efficiency, tend to command slightly higher rents. This suggests that the energy savings in green buildings could lead to increased rental income, supporting the case for green certification.
7. The plot shows an unclear relationship between gas costs and rent. However, if we observe the electricity cost and other utilities, we can see that utilities in general do have an impact on the rent being higher than usual.
8. A positive relationship between leasing rates and green certification suggests that green buildings might attract more tenants and lead to higher occupancy rates, which can stabilise revenue and improve ROI.
9. Investing in green certification for larger buildings might be more economically viable due to the potential for higher rent and occupancy rates.
10. Larger buildings might have more resources to invest in green certification, and the increased rent in larger buildings could make the return on investment for green certification more favourable.

CONCLUSION:

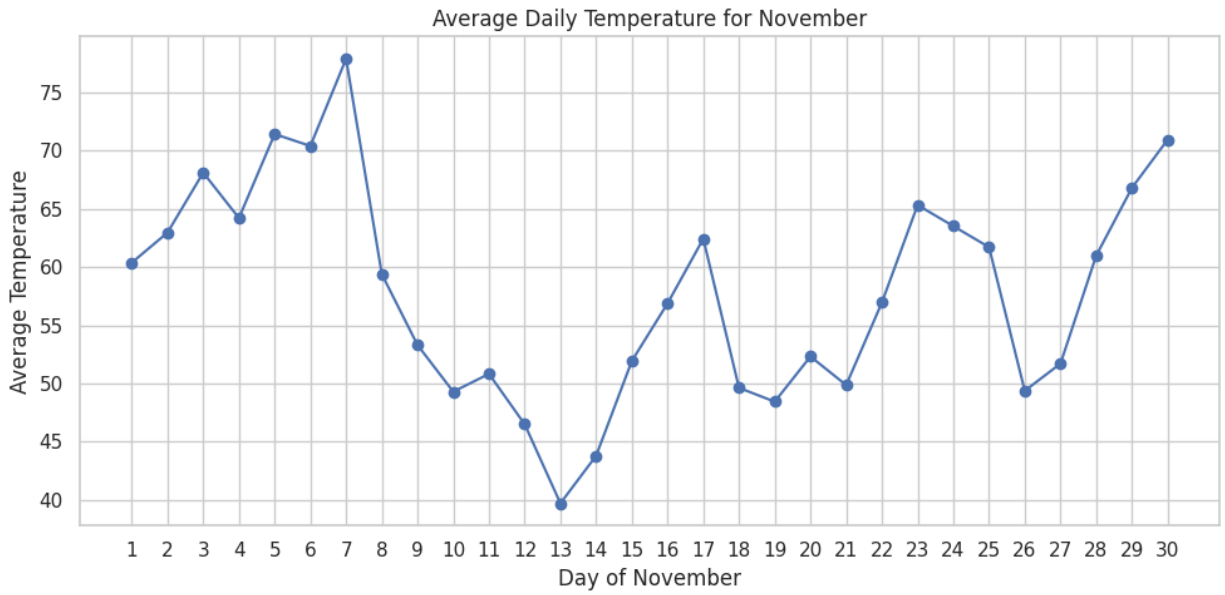
1. The analysis suggests that while green buildings might command higher rents and have better occupancy rates, the decision to invest in green certification should consider the size of the building, potential energy savings, and the local market's demand for sustainable spaces. Larger buildings, in particular, might see a more favourable return on investment due to these factors.
2. A more detailed regression analysis could quantify these relationships and help predict the economic returns from investing in green buildings. Additionally, a cost-benefit analysis that includes construction costs, potential rent premiums, and utility savings would provide a clearer picture of the financial viability of such an investment.

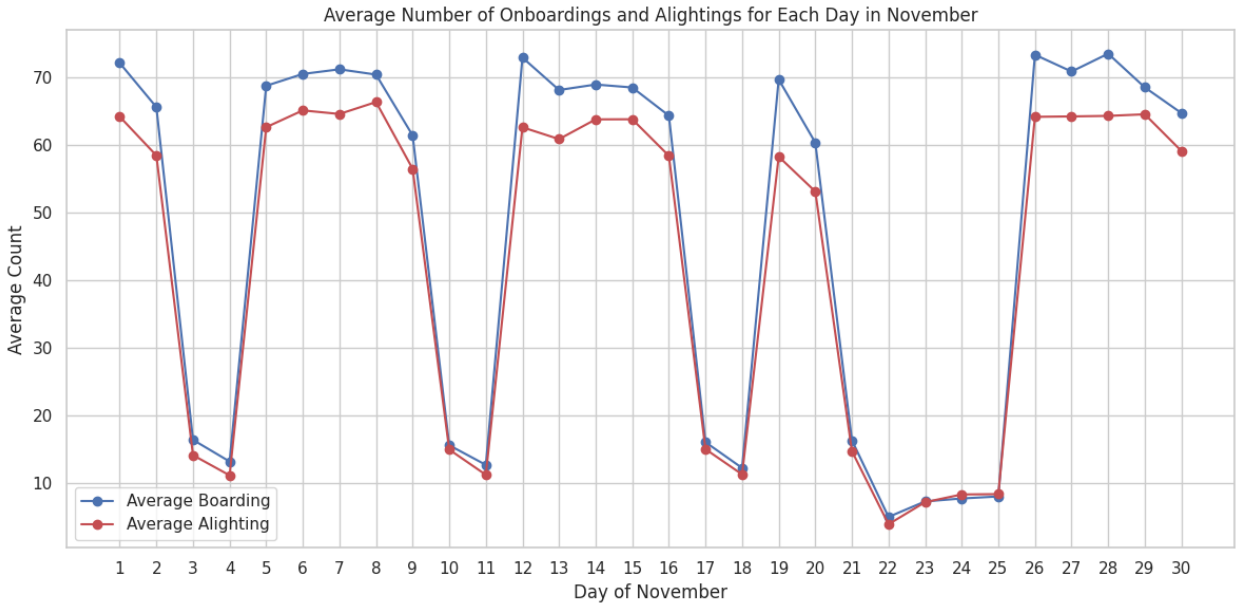
Q4. Visual Storytelling Part 2: Capital Metro data



- People take the Capital Metro more on weekdays than weekends
- There is a dip in the average number of people using the Capital metro, one of the reasons can be the near end of the Fall term and celebrations like Thanksgiving, another reason for the reduction can be seasonal as winter starts around November

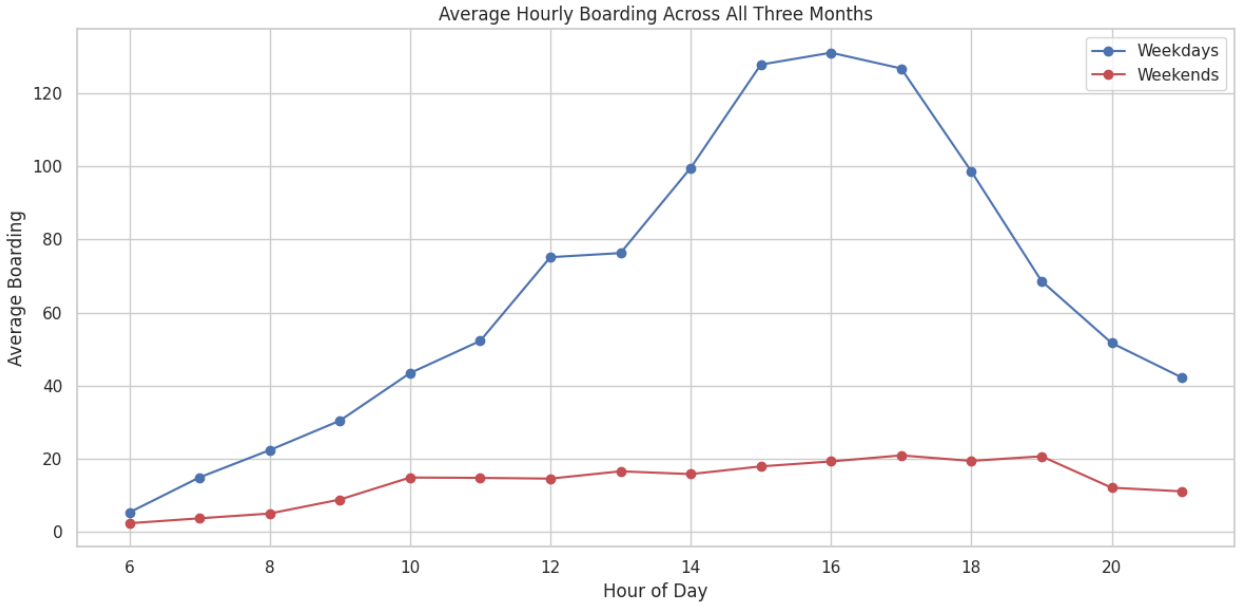
Let's explore the November being cold hypothesis and the Thanksgiving Hopethesis

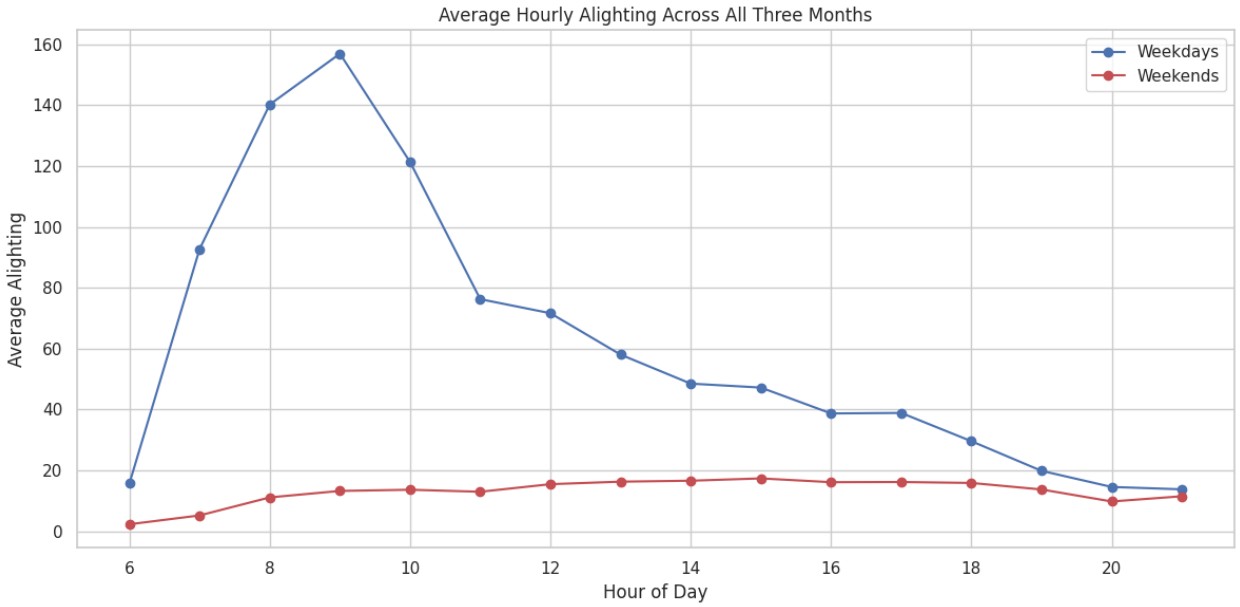




- Based on the graphs, it is evident that temperature does not significantly affect the number of people travelling via the Capital Metro. However, the Thanksgiving(22nd Nov, 2018) hypothesis appears to be accurate. We observe a substantial decrease in the number of passengers using the Capital Metro during this period, which

supports the assumption that people return home to celebrate Thanksgiving with their families.



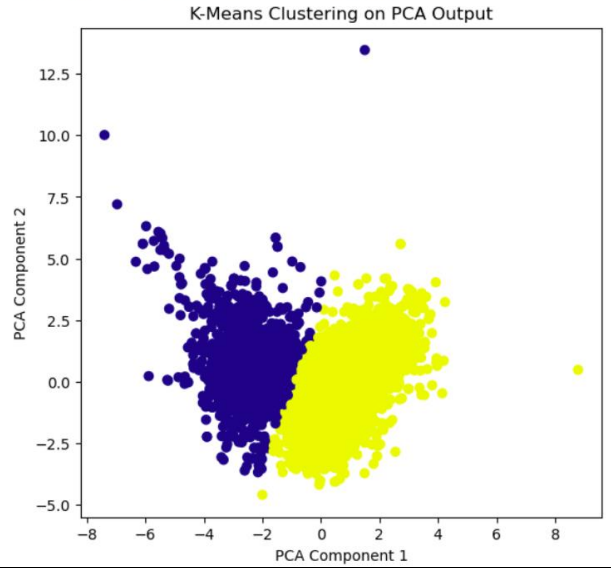


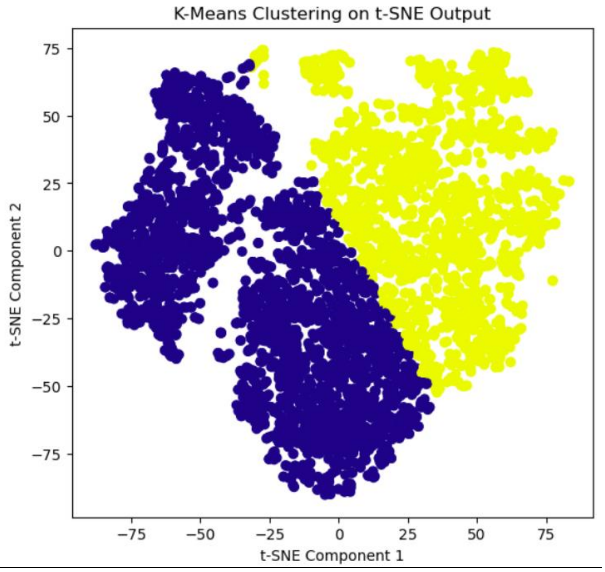
- The data shows that the average number of onboardings peaks between 3 to 5 PM, while the average number of alighting peaks between 8 to 10 AM.
- Given that the data includes shuttles to, from, and around the UT campus, it is logical that many passengers alight from the Capital Metro in the fall months when classes typically start between 8 to 11 AM and finish between 3 to 5 PM.

Q5. Clustering and dimensionality reduction

For wine color

```
Silhouette Score - Original Features: 0.2764785769202074
Silhouette Score - PCA Features: 0.46313927610201316
Silhouette Score - t-SNE Features: 0.3561696410179138
```

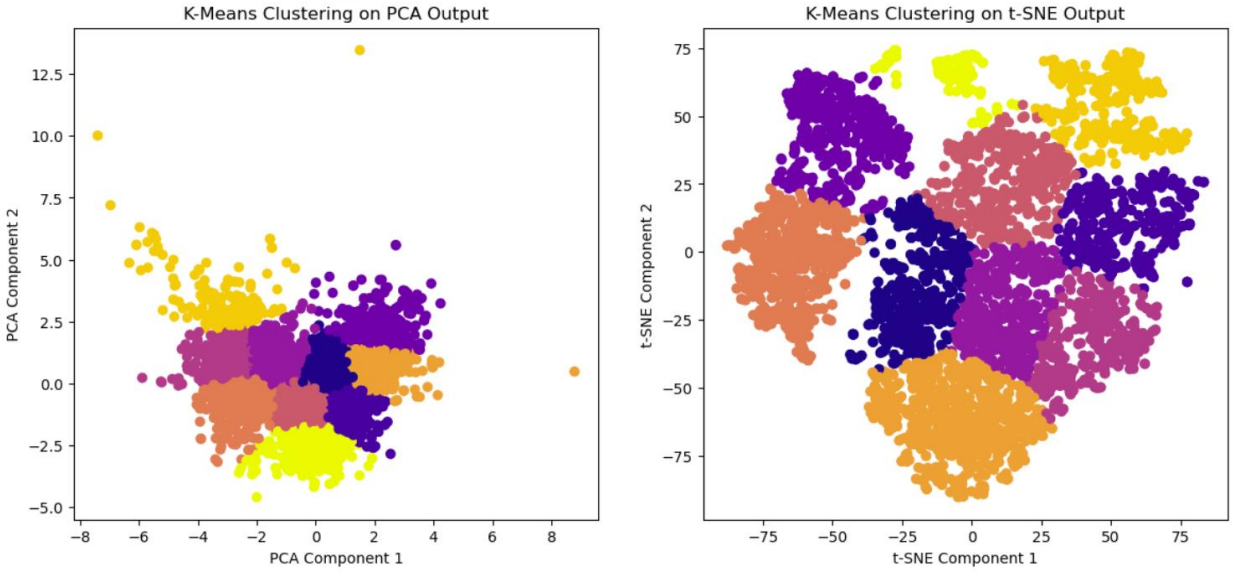




Based on the plots, compactness of clusters after PCA is higher than that after t-SNE, and the clusters formed in either case are not much separated from each other, which gives PCA the advantage. This is also reflected by the Silhouette Score, which is higher for PCA.

For wine quality

```
Silhouette Score - Original Features: 0.14492612500180016
Silhouette Score - PCA Features: 0.3439264448095275
Silhouette Score - t-SNE Features: 0.3971549868583679
```

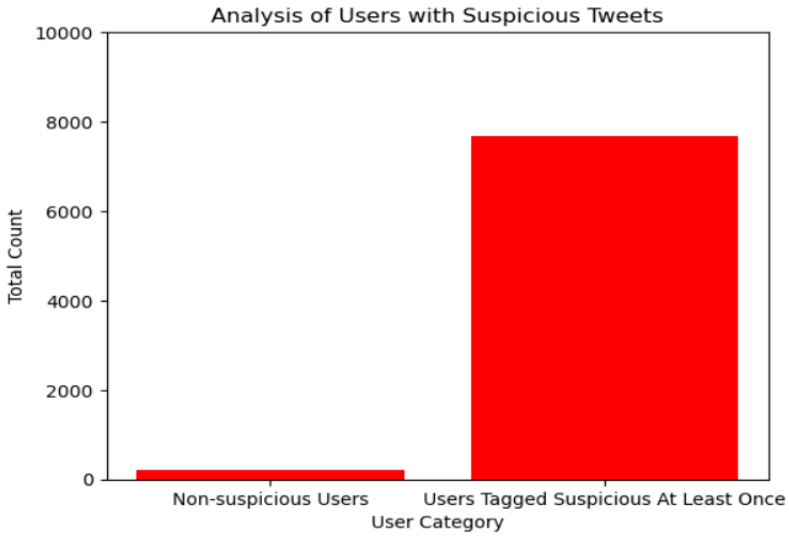


This time, the compactness of clusters in after PCA is again higher compared to t-SNE, but the separation among clusters is higher post t-SNE. This is why their Silhouette Scores are very close to each other and are low.

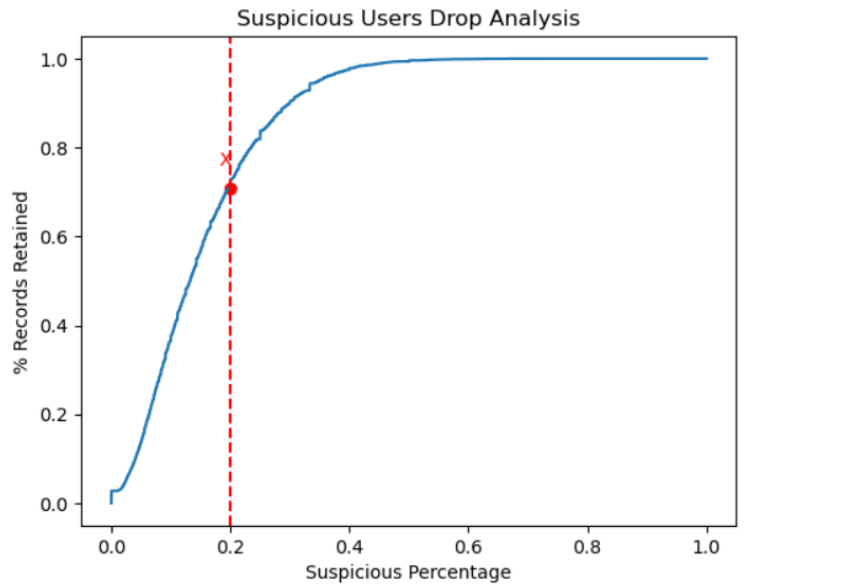
Q6) Market Segmentation

After loading the dataset we did the following steps for the preprocessing of the data:

- Rename the first column to "user_id"
- We can filter out non-useful tweets by removing users whose tweets fall under categories like chatter, spam, uncategorized, and adult, which we'll refer to as 'suspicious categories.'



- To refine this, we'll calculate a 'suspicious_percentage' metric, representing the proportion of a user's tweets labeled as suspicious, and use it to filter users effectively.
- The suspicious_category_counts variable is created by summing the values across all suspicious categories for each user. This gives a total count of suspicious activities per user.
- Similarly, total_category_counts calculates the total number of activities (across all categories) for each user.Finally, suspicious_percentage is calculated as the ratio of suspicious activities to the total activities for each user. This percentage indicates how much of a user's overall activity is deemed suspicious.

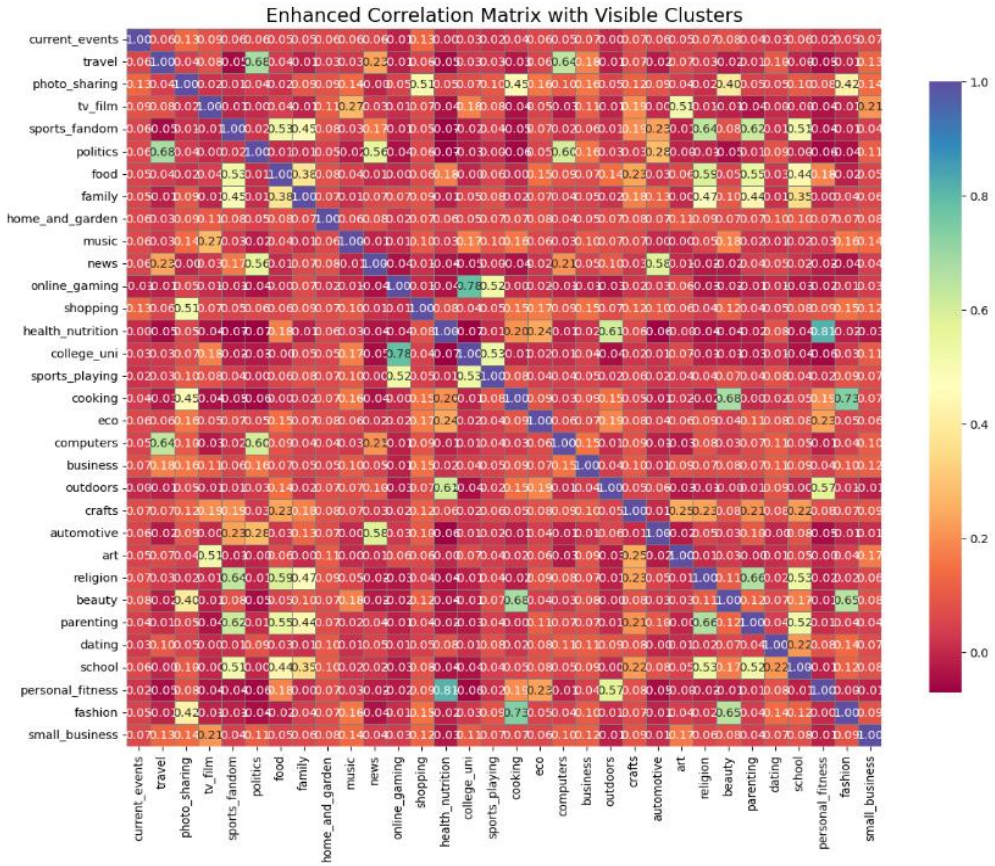


- We will set the filter threshold at 20%, meaning the users whose suspicious tweets is more than 20% of their total tweets will be filtered out from the dataset.Threshold is marked by X in above graph. Let us filter the data and remove these unwanted categories

  This filtering process was crucial to focus the analysis on users whose activities are less likely to be dominated by spam or irrelevant content. It helps in ensuring that the subsequent steps are based on more meaningful and reliable data.

Visualization Analysis

Let us try to plot a correlation matrix.



Immediately on a cursory glance, we can see some clusters in the plot above:

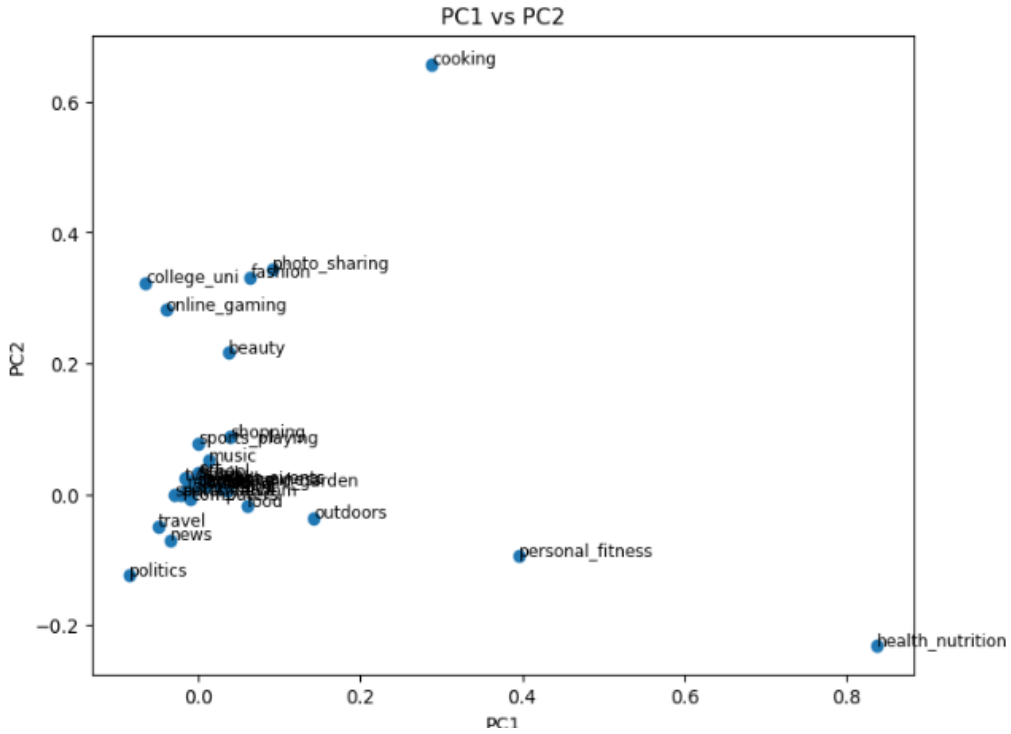
- **Lifestyle**: Strong correlations might exist between cooking, fashion, beauty, and home_and_garden.
- **Students and Gamers**: online_gaming,sports_playing,college_uni seem to correlate, reflecting student interests.
- **Cultural Interests:** art, tv_film,food,family,sports_fandom seem to be correlated due to a shared interest in entertainment and culture.
- **Political and News Engagement:** politics ,travel and news seem to be highly correlated, with potential links to current_events and travel.
- **Health and Fitness:** health_nutrition, personal_fitness, and cooking seem to correlate strongly within this group.

Let us try to reduce the dimensions using Principal Component Analysis

	Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
0	1	5.753916	0.221717	0.221717
1	2	4.477376	0.134252	0.355969
2	3	4.317003	0.124807	0.480776
3	4	4.199169	0.118086	0.598862
4	5	3.625283	0.088015	0.686877
5	6	2.447368	0.040112	0.726989
6	7	2.400296	0.038584	0.765572
7	8	2.245874	0.033779	0.799351
8	9	1.793203	0.021534	0.820886
9	10	1.599307	0.017129	0.838015

Inference: Clearly, we can observe that the proportion of variance explained increases very little after the addition of the 6th component. Thus, we will go ahead with 5 components and analyze them one by one

Now checking how are the individual PCs loaded on the original variables



PC1 helps outline health and personal care categories. Similarly analyzing PC2,PC3,PC4,PC5 and trying to visualize these Principal Components together in a bi-variate plot

Observations:

<u>PC4 vs. PC5:</u> This plot separates the family cluster from other clusters; the student cluster is also observed on the left side.

PC4 vs. PC2: This plot distinguishes the social media influencers cluster from the others.

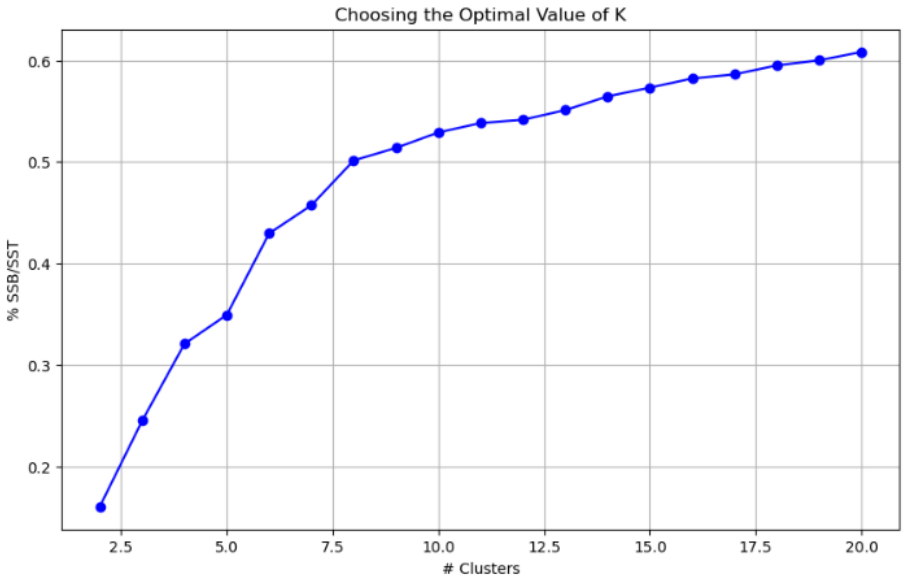
But still the inferences are not much clear. A more efficient approach maybe is to cluster these PCs and see which of them are similar

Final Conclusion of the PCA Analysis

We attempted to analyze all the principal components (PCs) to clearly identify the traits they distinguish, but we found it challenging to interpret them effectively. Since this is a marketing segmentation problem, having a clear understanding of what each PC represents is crucial. However, due to the lack of clarity in deciphering the PCs, we decided to abandon this approach and instead apply K-means++ clustering directly on the entire dataset.

K Means for Entire data

First, finding the optimum no of clusters



Along expected lines, **optimal number of clusters = 6**

Let us fit the final K means model to cluster it and form clusters

Clustering Insights.

Cluster 1: Home and Lifestyle Enthusiasts

Top Categories:

- Cooking: 374
- Fashion: 46
- Photo Sharing: 35

- Beauty: 9
- Sports Fandom: 7

Inference:

This cluster is predominantly composed of users interested in home and lifestyle topics. The strong focus on cooking suggests these users might be food enthusiasts or home cooks. The presence of fashion, beauty, and photo-sharing indicates a broader interest in lifestyle, personal appearance, and sharing their experiences online.

Cluster 2: Students and Gamers

Top Categories:

- Online Gaming: 191
- College/University: 183
- Sports Fandom: 4
- Health/Nutrition: 2
- Religion: 2

Inference:

This cluster seems to be dominated by students and gamers. The significant presence of online gaming and college/university categories indicates that these users are likely younger, possibly students with a strong interest in gaming. The lower counts in sports fandom, health/nutrition, and religion suggest these are secondary interests.

Cluster 3: Cultural and Artistic Enthusiasts

Top Categories:

- Sports Fandom: 220
- Religion: 188
- Art: 182
- TV/Film: 180
- Photo Sharing: 179

Inference:

This cluster appears to consist of users with a strong inclination toward cultural and artistic pursuits. Sports fandom is the most prominent, but the high numbers in art, religion, TV/film, and photo-sharing suggest these users have diverse

interests in cultural activities and likely enjoy engaging with content related to entertainment, art, and spirituality.

Cluster 4: Politically Engaged Travelers

Top Categories:

- Politics: 212
- Travel: 113
- Photo Sharing: 7
- News: 4
- Art: 3

Inference:

Users in this cluster are likely to be politically aware and engaged, with a strong interest in travel. The high record count in politics suggests that these users frequently engage with political content. Their interest in travel and photo-sharing indicates that they enjoy exploring new places and sharing their experiences, possibly reflecting a global outlook.

Cluster 5: News and Automotive Enthusiasts

Top Categories:

- News: 214
- Politics: 96
- Automotive: 60
- Sports Fandom: 30
- Photo Sharing: 11

Inference:

This cluster is characterized by a strong interest in news and politics, with a significant segment also focused on automotive topics. Users in this cluster likely keep themselves informed about current events and enjoy discussions about politics. The presence of automotive as a top category suggests an additional interest in cars and possibly other vehicles.

Cluster 6: Health and Fitness Enthusiasts

Top Categories:

- Health/Nutrition: 739
- Personal Fitness: 87
- Dating: 18
- Photo Sharing: 17

- Art: 16

Inference:

This cluster is dominated by users who are highly focused on health and fitness. The overwhelming presence of health/nutrition and personal fitness as top categories indicates that these users are likely fitness enthusiasts, possibly interested in maintaining a healthy lifestyle. The smaller interests in dating, photo-sharing, and art suggest that these users also engage in social interactions and appreciate creative content, but these are secondary to their primary focus on health.

Summary of Cluster Characteristics:

Cluster 1: Home and Lifestyle Enthusiasts - Focused on cooking, fashion, and beauty.
Cluster 2: Students and Gamers - Dominated by online gaming and college-related content.
Cluster 3: Cultural and Artistic Enthusiasts - Interested in sports, art, religion, and entertainment.
Cluster 4: Politically Engaged Travelers - Focused on politics and travel, with a global outlook.
Cluster 5: News and Automotive Enthusiasts - Interested in news, politics, and automotive content.
Cluster 6: Health and Fitness Enthusiasts - Strongly focused on health, nutrition, and fitness.

So, clearly these clusters are consistent with what we observed in the correlation plot during our EDA.

Q7) The Reuters corpus

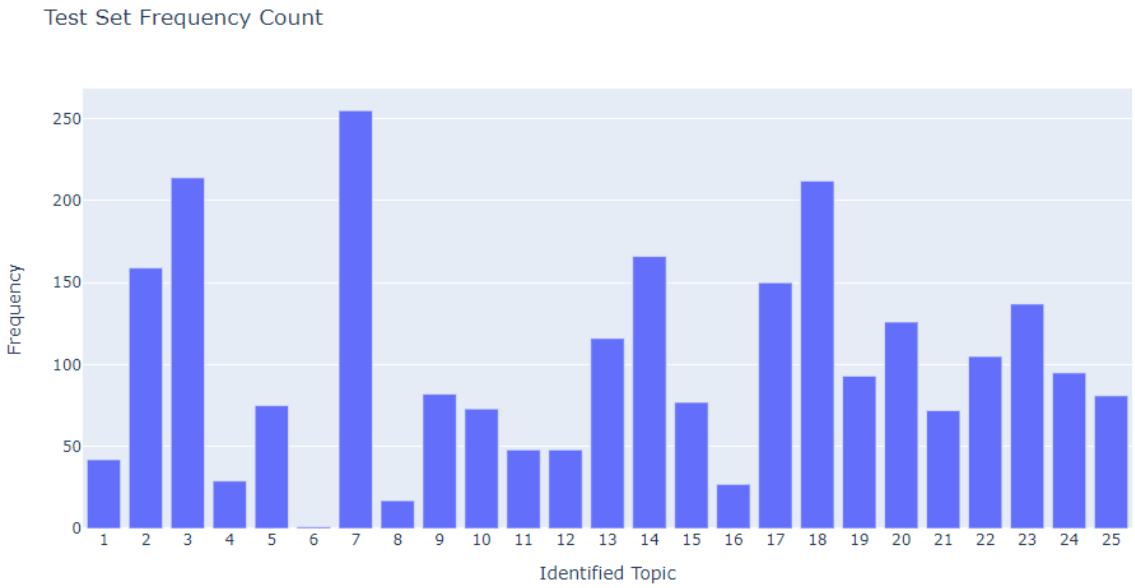
- **Question:** Identify topics from the text data and summarize your findings.
- **Approach**: We applied **Latent Dirichlet Allocation (LDA),** a generative probabilistic model, to uncover key topics within the training data. LDA operates under the assumption that each document in the corpus is composed of a limited set of topics, with each word linked to one of these topics. Subsequently, we identified the most dominant topic within the test set files.
- **Results**: Some of the topics that were identified through LDA were:

Topic	Keywords	Class
Topic 1	united, states, trade, said, drug, china, ban, department, colombia, congress	**International Trade**
Topic 2	hong, kong, china, said, tung, chinese, people, territory, rule, says	**Hong Kong-China Relations**
Topic 3	internet, corp, new, computer, said, software, technology, microsoft, network, services	**Technology and Software**
Topic 4	said, financial, chairman, president, statement, company, vice, board, right, street	**Corporate Leadership and Governance**
Topic 5	amp, local, long, market, competition, service, phone, cable, rules, companies	**Telecommunications and Market Competition**
Topic 6	told, reuters, director, interview, reporters, quality, telephone, areas, conference, managing	**Media and Communication**
Topic 7	china, said, beijing, chinese, official, taiwan, officials, economic, communist, state	**Chinese Economy and Policy**
Topic 8	news, said, early, fund, 1997, joint, year, venture, start, 1998	**News and Media Industry**
Topic 9	000, tonnes, said, saying, 100, cocoa, year, copper, 500, figures	**Commodity Markets**

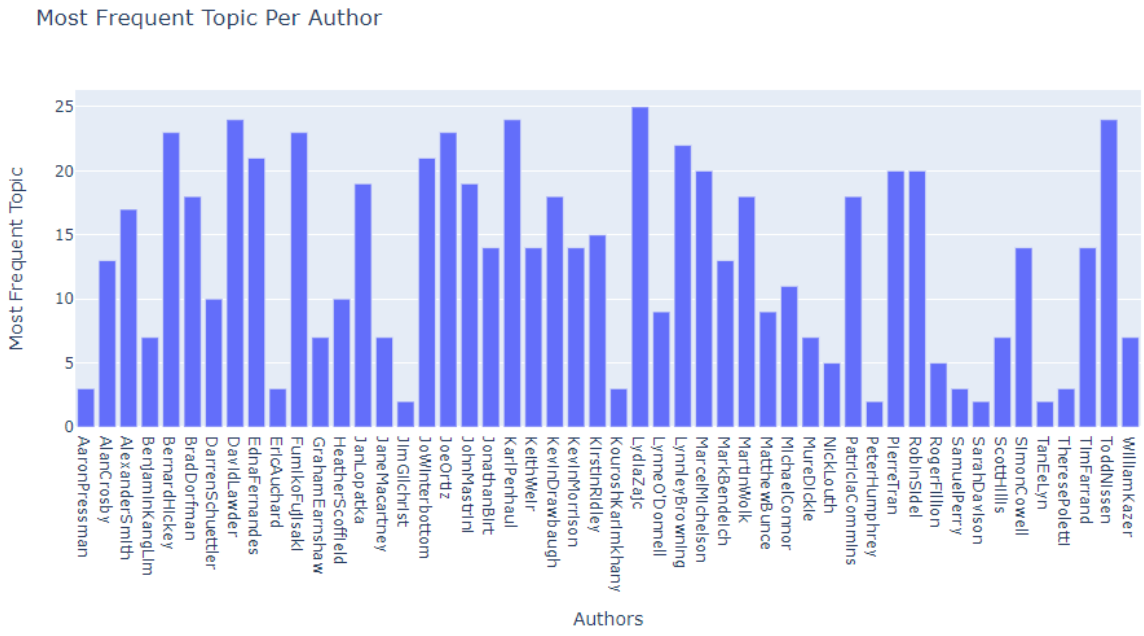
Topic 10	percent, gold, price, said, share, market, 20, bre, 15, stocks	**Financial Markets and Investment**
Topic 11	said, wang, court, rights, case, chinese, given, government, details, action	**Human Rights and Legal Cases**
Topic 12	chief, executive, said, company, years, officer, new, ago, development, chairman	**Business Leadership and Development**
Topic 13	bank, banks, year, percent, rate, canada, central, credit, said, banking	**Banking and Finance**
Topic 14	million, year, pounds, profit, profits, half, 1995, net, percent, 30	**Financial Performance**
Topic 15	deal, company, largest, merger, world, mci, bt, british, stake, percent	**Mergers and Acquisitions**
Topic 16	billion, year, total, debt, said, francs, worth, ve, percent, got	**Financial Transactions and Debt**
Topic 17	said, analyst, think, going, market, term, good, don, people, added	**Market Analysis and Forecasting**
Topic 18	quarter, sales, year, said, earnings, share, analysts, expected, results, company	**Corporate Earnings and Results**
Topic 19	government, said, czech, general, minister, ahead, finance, party, house, ministry	**Government and Politics**
Topic 20	said, offer, shareholders, french, american, airbus, south, company, north, new	**Corporate Shares and Ownership**
Topic 21	british, pence, bid, air, share, plc, group, said, dividend, britain	**Stock Market and Investments**
Topic 22	said, prices, oil, year, demand, russia, domestic, set, rates, export	**Oil and Energy Markets**
Topic 23	business, said, life, insurance, japan, market, financial, non, group, banks	**Business and Finance in Japan**
Topic 24	gm, said, workers, plant, agreement, union, ford, car, plants, comment	**Automotive Industry and Labor Relations**

Topic 25	stock, shares, trading, exchange, new, close, york, points, share, toronto	**Stock Market Trading and Exchanges**

After analyzing the test data files, it was found that topic 7 is the most prevalent topic in the test files.



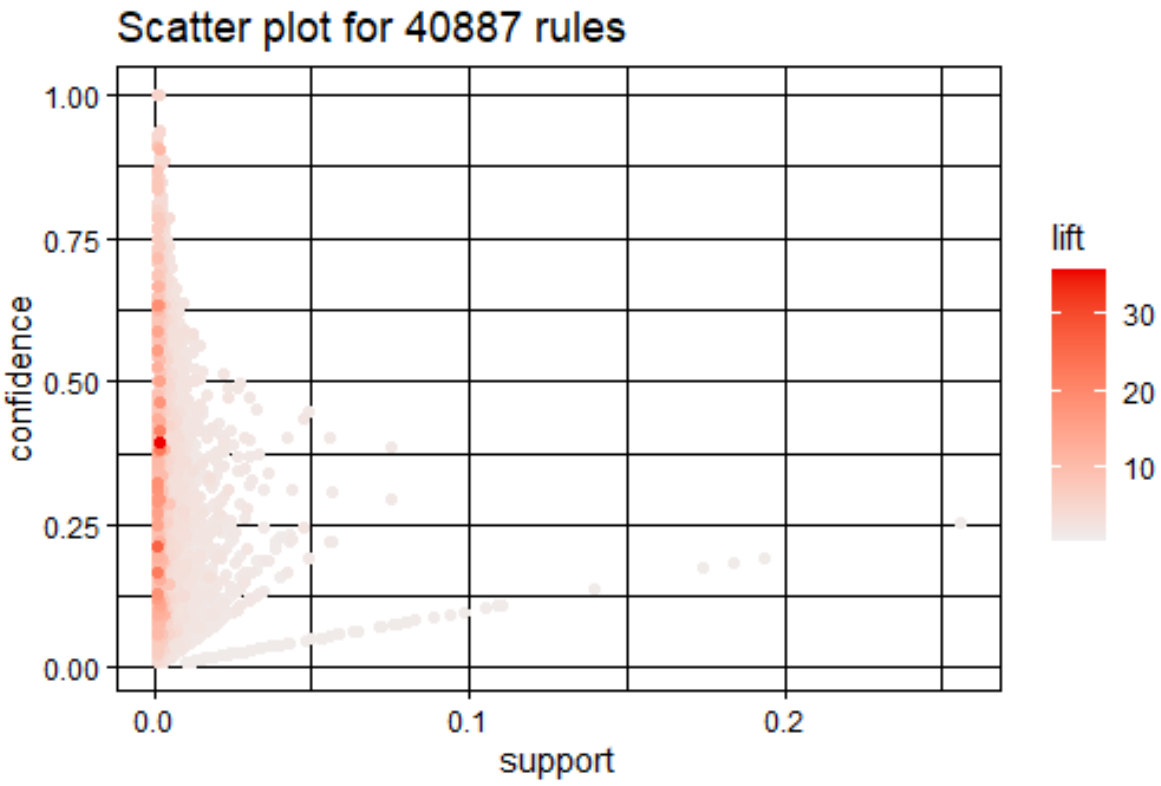
We also have a summary of the most frequent topics written by each of the writers.



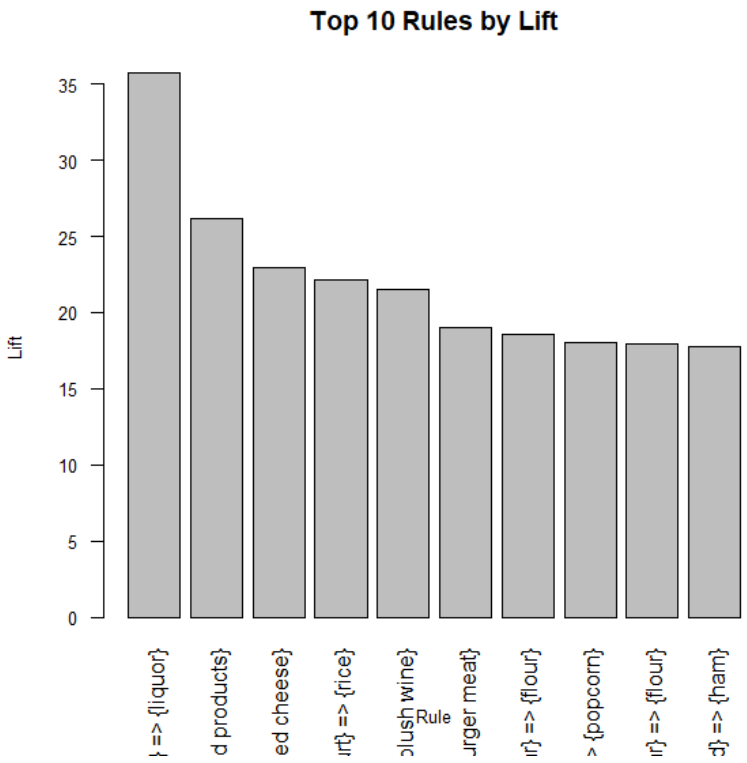
- **Conclusion:** We identified distinct topics, each characterized by a set of keywords that provided meaningful context. The results of our analysis showcased a diverse range of topics present in the dataset, from global trade and technology advancements to legal cases and economic developments.s. We conducted an analysis of the test data files and determined that topic 7, centered around **economic and geopolitical issues in China,** emerged as the most prevalent topic among the test articles. This finding emphasizes the significance of this particular theme within the writings of the authors, potentially highlighting their expertise and areas of interest.

Q8) Association rule mining

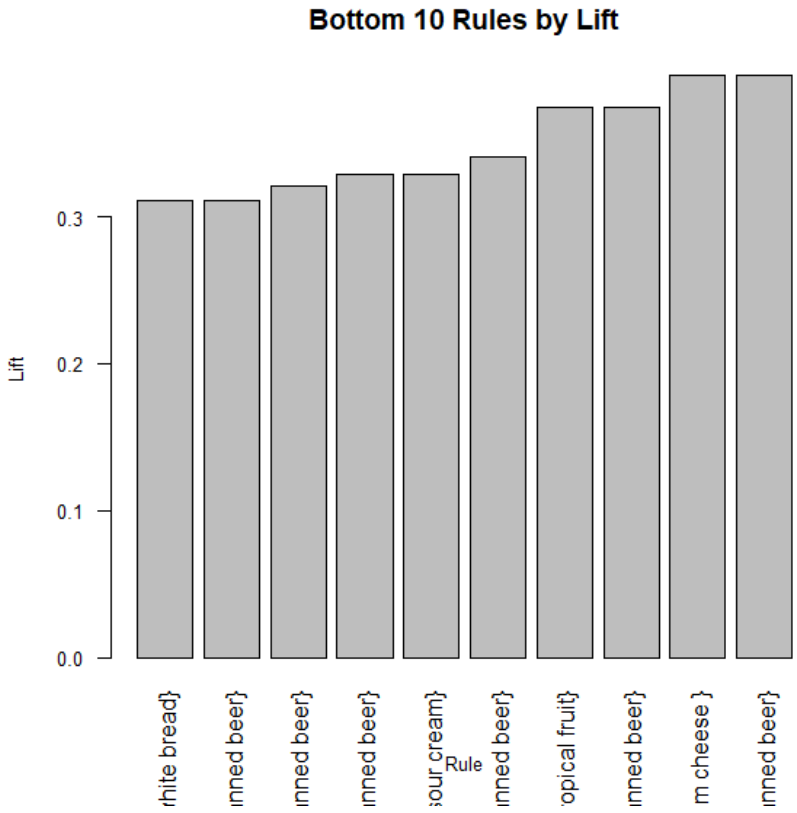
We chose the max_length of 9 and set the confidence interval at 0.6 and the lift parameter greateer than 10.
The choice was made as both the values are on healthy note, and since our max length parameter is relatively high, it is bound to do better hence large values of lift and confidence is selected.
Rather than choosing a random support parameter we iterated over the range of values and then chose the parameters that provided a higher lift value, following is the scatter plot confidence interval vs support where the red color gives the lift value.



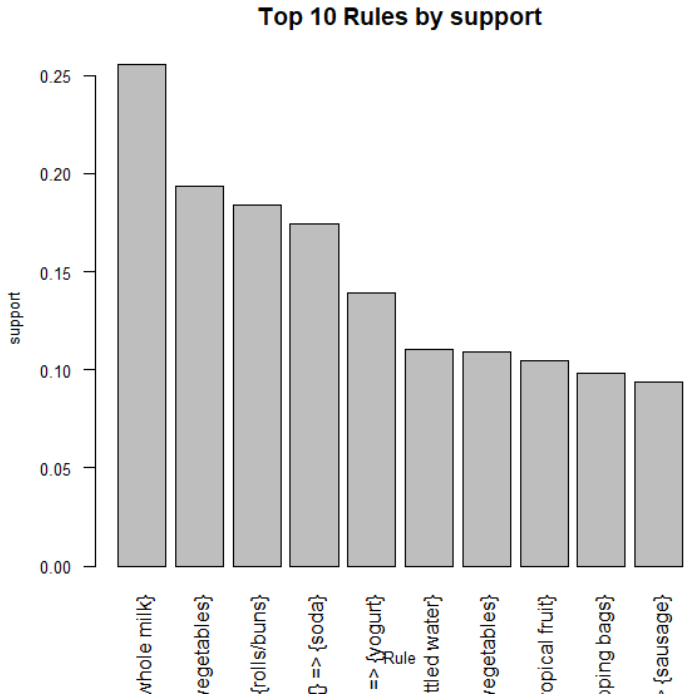
Following are the best and worst lift values combinations for the decided parameters. The top lift rules, suggests high likelihood of similar items. For example, we can categorize cart broadly as dairy products, liquor cart's suggest party preparation, and also confectionary ingredients seems to be clubbed together quite often.



The bottom 10 lift's sets show relatively low confidence values, suggesting vague correlation, and support values are also very low, indicating that the business might profit if these are offered/purchased independently.



Further we also observed from the top support plot that each category there was on their own. The items were as expected common items like milk, soda, vegetables and bread.



Q9) Image classification with neural networks

Model Architecture:

- The neural network consists of 3 convolution 2d layers, 1 max pool layer and 2 linear layers, including fully connected layers. The specific architecture details (e.g., number of layers, neurons per layer) are defined in the model class, which is designed for the classification task.

Hyperparameters:

- **Batch Size:** The batch size is specified to be 100, determining the number of training examples utilized in one iteration before updating the model's parameters.
- **Number of Epochs:** The number of epochs is defined to be 5, indicating how many times the entire dataset will be passed through the network during training.

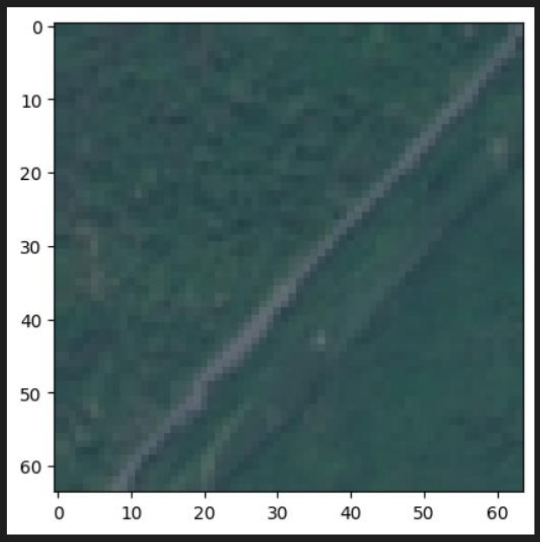
Training Process:

- **Loss Function:** The loss function used is Cross-Entropy Loss, which is typical for classification tasks.
- **Optimizer:** The optimizer Adam is utilized with its parameters configured to control the weight update mechanism during training.

Performance Metrics:

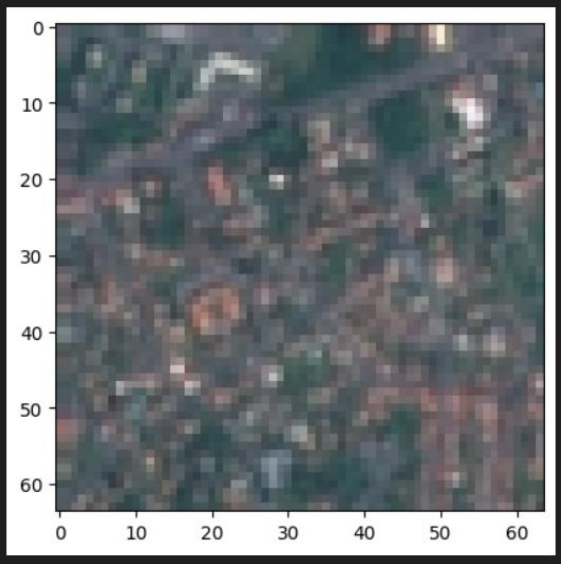
- **Accuracy:** The model's accuracy on the test dataset is calculated and displayed, providing a measure of the model's performance. The accuracy was found out to be **0.8752**.
- **Confusion Matrix:** A confusion matrix with dimensions corresponding to the number of classes (e.g., 10 classes) is generated to visualize the classification results, showing true positives, false positives, false negatives, and true negatives for each class.

```
Predicted: 1
Actual: 1
```

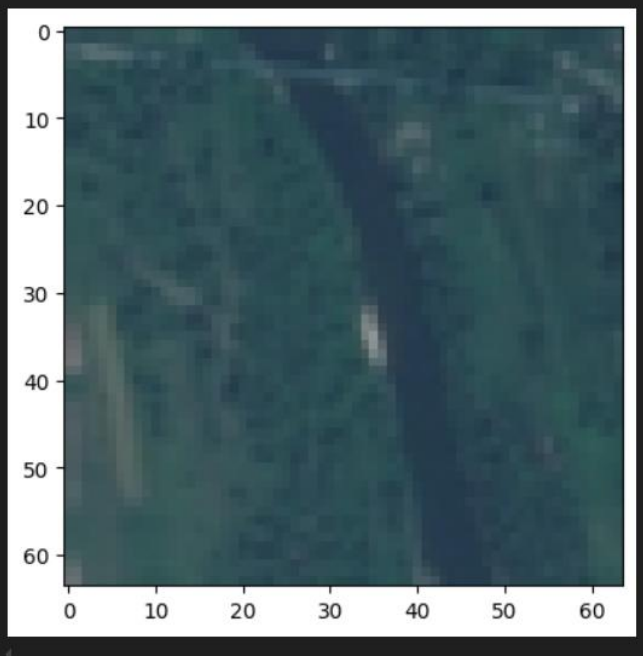



```
Predicted: 3
Actual: 3
Epoch 5, Loss: 0.2890144807435352
Finished Training
```

```
Predicted: 5
Actual: 8
```

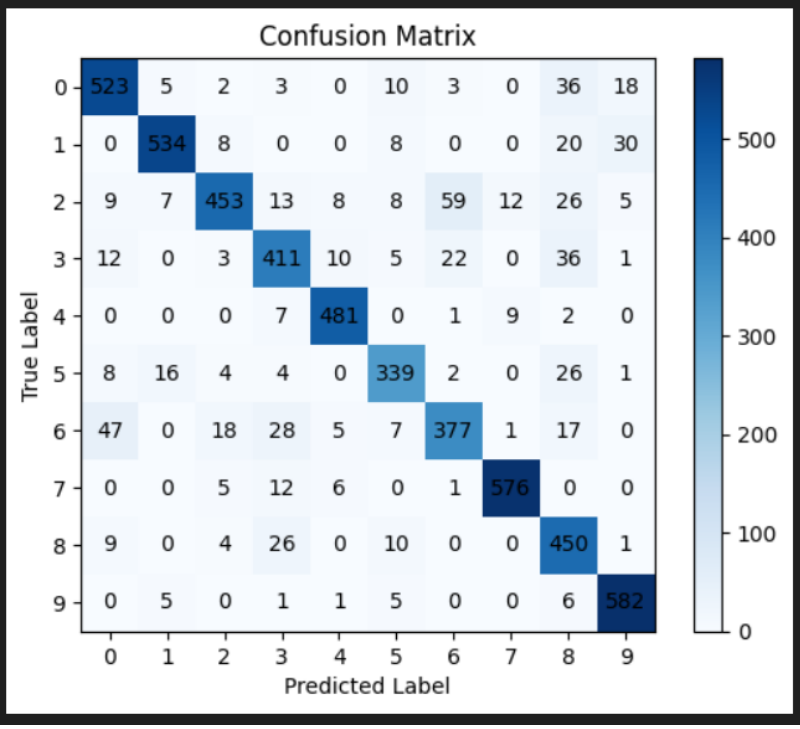


```
Predicted: 3
Actual: 3
Epoch 4, Loss: 0.3564061505290178
```



The model accuracy is 87.52% and below is confusion matrix

```
Finished Training: Total Correct: 4726/5400 (0.8752 Accuracy)
```



The true label signifies the original output while the predicted label explain is the values predicted by the NN architecture created by us.