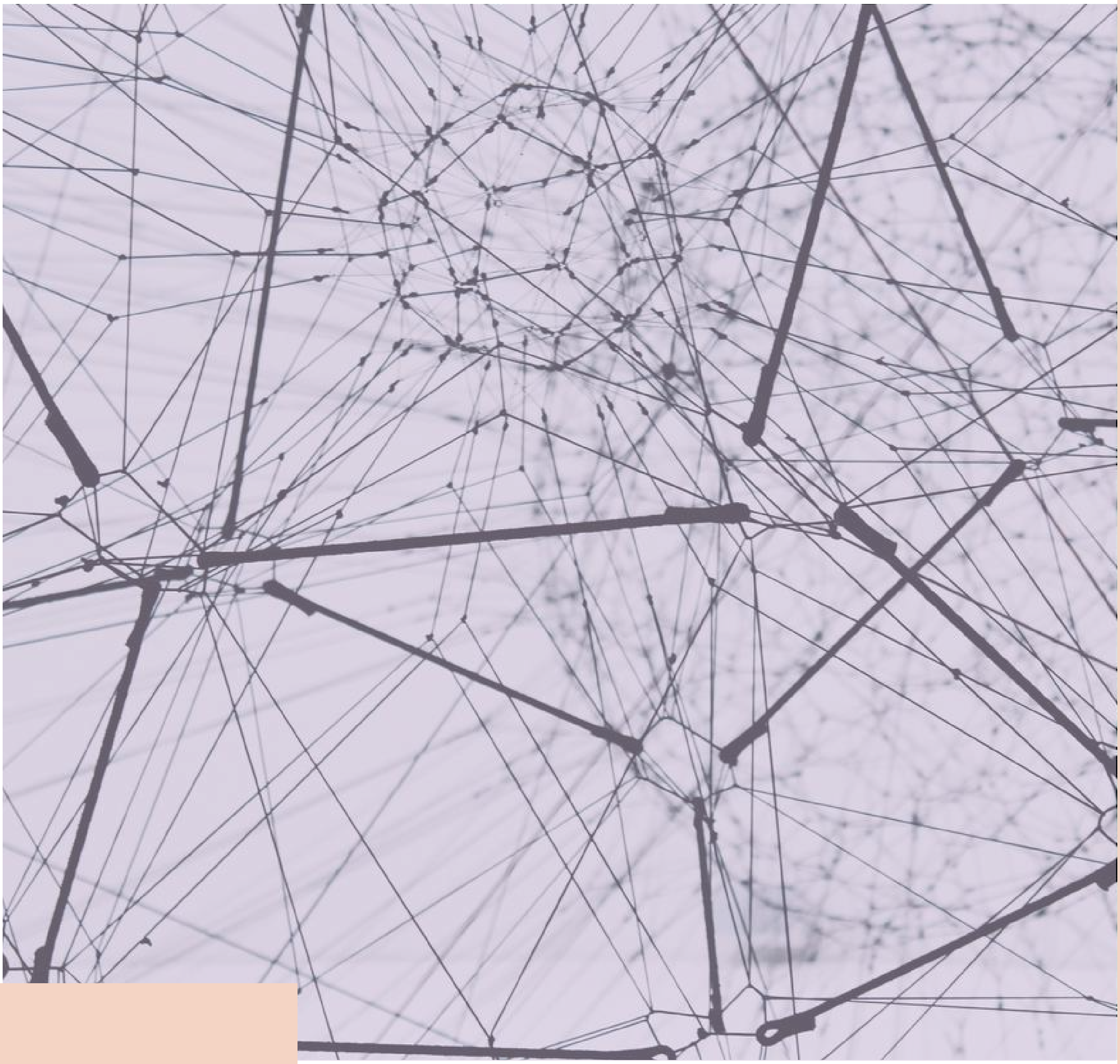


DATA SCIENCE PORTFOLIO

NICKEN SHIDQIA NURAHMAN

Selected Work
2023-2024

Nicken Shidqia Nurahman

As a recent Civil Engineering graduate with 2 years of experience in Construction Management, I bring a strong foundation in data analysis, SQL, Python, and Tableau. Eager to apply my analytical skills and practical knowledge to excel in a Data Scientist or Data Analyst role.

081389405425

nicken.shidqia@yahoo.co.id

https://www.linkedin.com/in/nickenshidqia/

https://github.com/nickenshidqia/Data_Science_Portfolio

EDUCATION

Universitas Indonesia

2021 - 2023

Masters Degree in Project Management, 3.92/4.00

Universitas Trisakti

2015 - 2019

Bachelor Degree in Civil Engineering, 3.51/4.00

SKILLS & CERTIFICATION

- Data Science Bootcamp - Rakamin Academy (6 Months) (2024)
- The Data Science Course: Complete Data Science Bootcamp 2023 - Udemy (2023)
- The Complete SQL Masterclass 2023 - Udemy (2023)
- Python Mega Course: Learn Python - Udemy (2023)
- Tableau Data Analyst Certification Prep 2024 - Udemy (2024)

WORK EXPERIENCES

Project-Based Virtual Intern : Big Data Analytics Kimia Farma x Rakamin Academy

Jan 2024 - Feb 2024

Data Analyst Intern

Project-Based Virtual Intern : Data Scientist Home Credit Indonesia x Rakamin Academy

Dec 2023 - Jan 2024

Data Scientist Intern

Project-Based Virtual Intern : Data Scientist ID/X Partners x Rakamin Academy

Nov 2023 - Dec 2023

Data Scientist Intern

Project-Based Virtual Intern : Data Scientist Kalbe Nutritionals x Rakamin Academy

Oct 2023 - Nov 2023

Data Scientist Intern

PT. ISTAKA KARYA (Persero)

Aug 2019 - Sep 2021

Engineering Administration and Project Control Staff

PT. ISTAKA KARYA (Persero)

Feb 2019 - Jul 2019

Project Control Intern Staff

TABLE OF CONTENTS

MACHINE LEARNING PROJECT

SQL Project, Data & Dashboard Analysis

Machine Learning Project (Regression & Clustering) on Kalbe Nutritionals

Click here to get full code

Project Description

Data scientist in Kalbe Nutritionals got a new project from :

1. Inventory Team to predict sum of quantity from all products, so they could create sufficient daily inventory.
2. Marketing Team to create cluster or segment of customer to get personalized promotion and sales treatment

Project Result

Data Ingestion and Exploratory Data Analysis using PosgreSQL & DBeaver

Query 1 : Average customer age based ontheir marital status

Marital Status	age_average
	31.3333333333
Married	43.0382352941
Single	29.3846153846

Query 2 : Average customer age based on their gender

gender	age_average
Wanita	40.326446281
Pria	39.1414634146

Query 3 : Store name with the highest total quantity

storename	total_quantity
Lingga	2,777

Query 4 : The best-selling product with the highest total amount

Product Name	total_amount
Cheese Stick	27,615,000

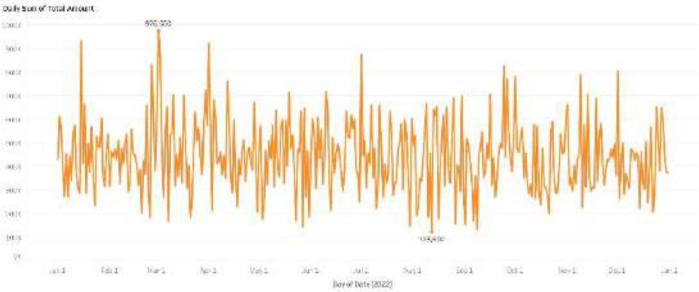
Dashboard Visualization Using Tableau

Worksheet 1 : Total quantity from month to month

Insight :

- Sales trends fluctuate slightly, and show a gradual decline starting from June.
- The highest total quantity sold is in March 2022 with 1,753 items
- The lowest total quantity sold is in December 2022 with 1,409 items

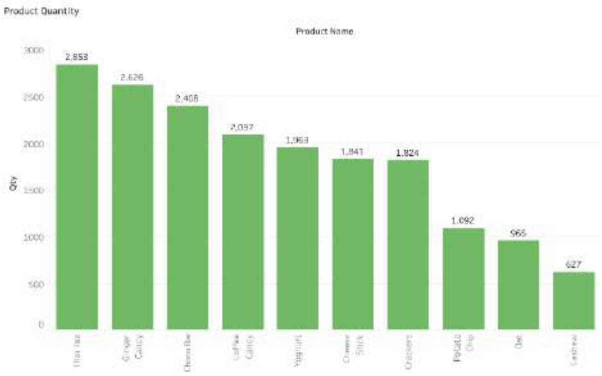
Worksheet 2 : Total amount from day to day



Insight :

- Daily revenue trends fluctuate heavily
- The highest total amount is in March 2022 with Rp 976,500
- The lowest total amount is in August 2022 with Rp 123,600

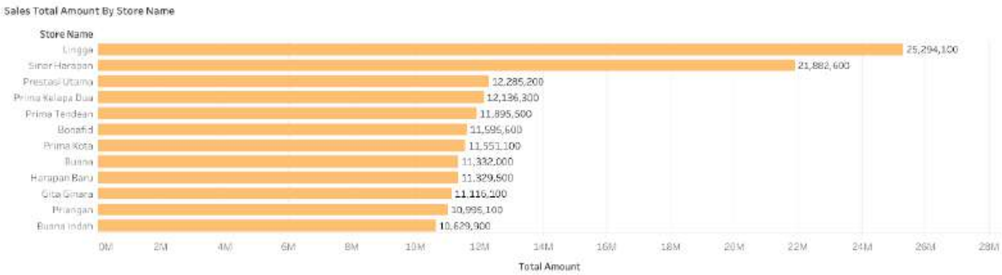
Worksheet 3 : Total quantity by product



Insight :

- The highest selling product in 2022 is Thai Tea with 2,853 items sold.
- The lowest selling product in 2022 is Cashew with 627 items sold.

Worksheet 4 : Total sales amount by store name



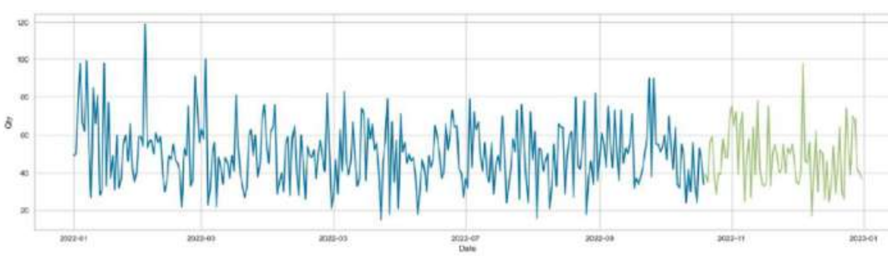
Insight :

- The best-selling store in 2022 is Lingga with sales revenue reached Rp 25,294,100.
- The lowest-selling store in 2022 is Buana Indah with sales revenue reached Rp 10,629,900.

Daily Product Quantity Prediction Using Time Series Arima

Data Training & Testing

Splitting the data with 80% training and 20% testing. Blue line is data training, and green line is data testing.



Find p,d,q for ARIMA Model

Model 1 - Auto-fit ARIMA

Get result with p,d,q = 1,0,1. ARIMA (1,0,1) means there is no Differencing (0) because it is stationary, with Autoregression for 1 lag and 1 order Moving Average.

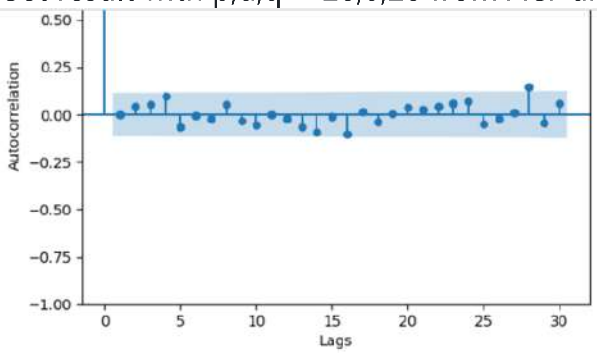
Dep. Variable:	y	**No. Observations:**	292
Model:	SARIMAX(1, 0, 1)	**Log Likelihood**	-1244.943
Date:	Mon, 23 Oct 2023	**AIC**	2495.886
Time:	06:09:33	**BIC**	2506.916
Sample:	01-01-2022	**HQIC**	2500.304
	- 10-19-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0000	4.8e-05	2.08e+04	0.000	1.000	1.000
ma.L1	-0.9830	0.016	-60.722	0.000	-1.015	-0.951
sigma2	290.0520	24.022	12.074	0.000	242.969	337.135

Ljung-Box (L1) (Q):	0.11	**Jarque-Bera (JB):**	8.00
Prob(Q):	0.74	**Prob(JB):**	0.02
Heteroskedasticity (H):	0.70	**Skew:**	0.39
Prob(H) (two-sided):	0.08	**Kurtosis:**	3.24

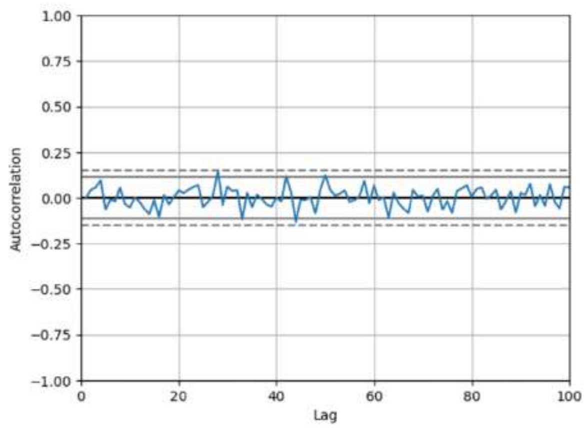
Model 2 - ACF & PACF Plot

Get result with p,d,q = 28,0,28 from ACF and PACF plot.



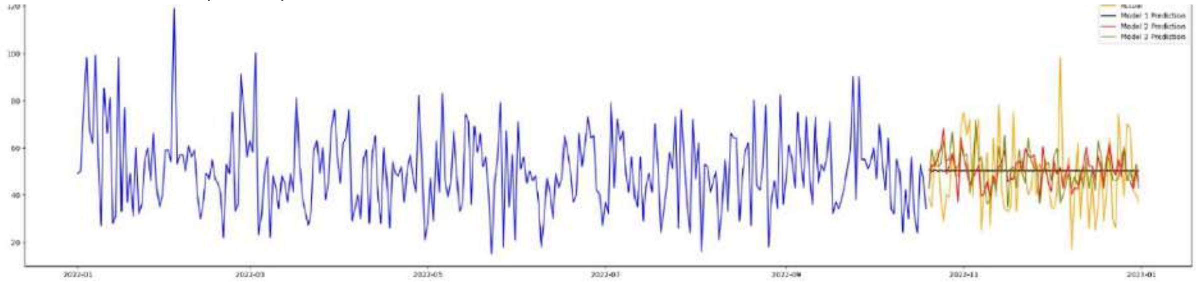
Model 3 - Autocorrelation Plot

Get result with p,d,q = 44,0,44 from Autocorrelation plot.



ARIMA Modelling Plot

Plot Data Train, Test, and Model Prediction



Forecast Quantity Sales With The Best Parameter

Model 2 with p,d,q (28,0,28) show the best metric evaluation because has the lowest MAE, MSE, RMSE, and MAPE.

```
Model 2
Mean Absolute Error (MAE) = 13.10
Mean Squared Error (MSE) = 255.45
Root Mean Squared Error (RMSE) = 15.98
Mean Absolute Percentage Error (MAPE) = 31.88%
```

Prediction for quantity on January 2023 is 50 pcs/day

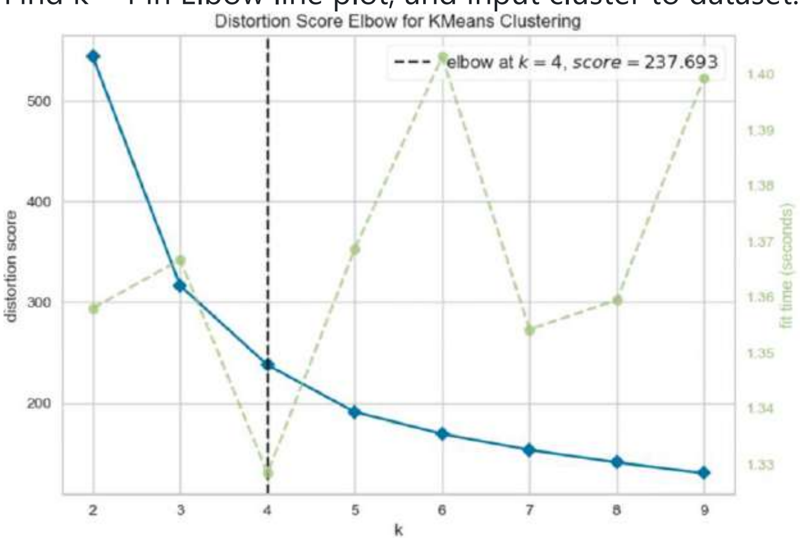
```
df_future_model2_out.describe()
```

```
count    30.000000
mean     50.116869
std       7.303344
min      37.041028
25%      45.600319
50%      49.702099
75%      53.784914
max      68.038147
Name: Predictions, dtype: float64
```

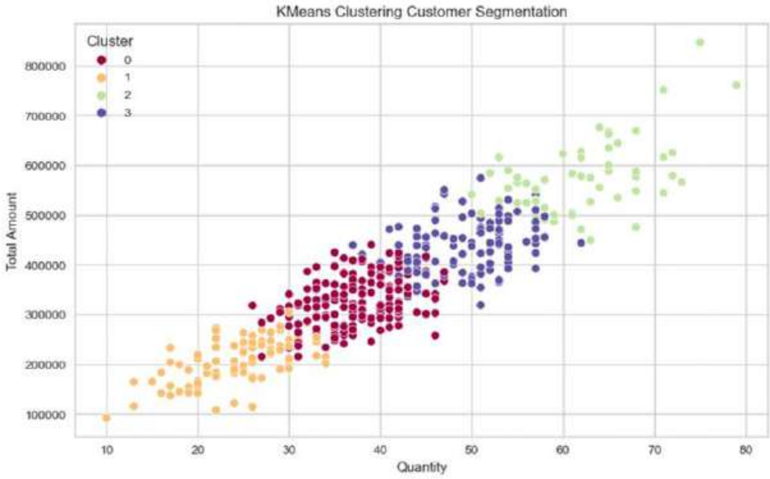
Customer Segmentation Using KMeans Clustering

Elbow Method

Find k =4 in Elbow line plot, and input cluster to dataset.



There are 4 cluster of segmentation :



Conclusion

- Cluster 0 is the cluster with the most largest number of customers, but has the second lowest average of quantity and total amount. The strategy is give special offering and discount for new member
- Cluster 1 is the cluster with the second fewest number of customers, and the lowest average of quantity and total amount. One of the strategy is collaborate with influencers to promote products.
- Cluster 2 is the customer that valuable to the business. the strategy is offer loyalty membership
- Cluster 3 has the second largest average of quantity and total amount and has potential of upselling.

Machine Learning Project Credit Risk Assessment (Using Logistic Regression) on ID/X Partners

Project Description

A primary risk with corporate loans is failing in accurately assessing credit risk. Disadvantage of Manual credit risk assessment :

- Subjectivity can introduce bias and inconsistency in decision-making
- Time-consuming especially when dealing with a large number of loan applications.
- Humans errors, such as data entry mistakes, miscalculations, or oversight of important details

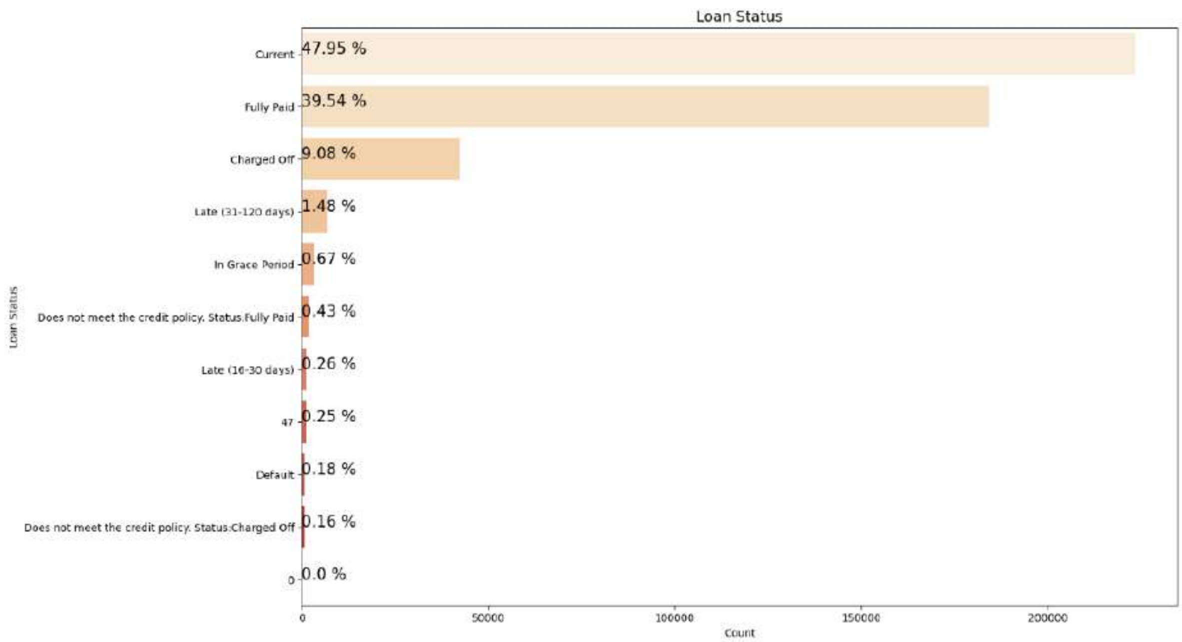
Challenges :
Build a machine learning model that can predict credit risk assessment

Project Result

[Click here to get full code]

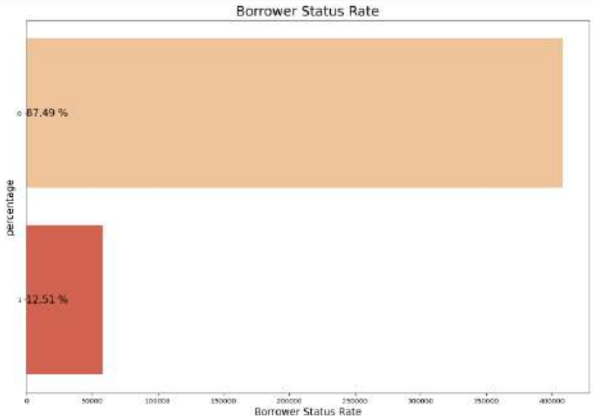
Exploratory Data Analysis

Applicants by Loan Status :



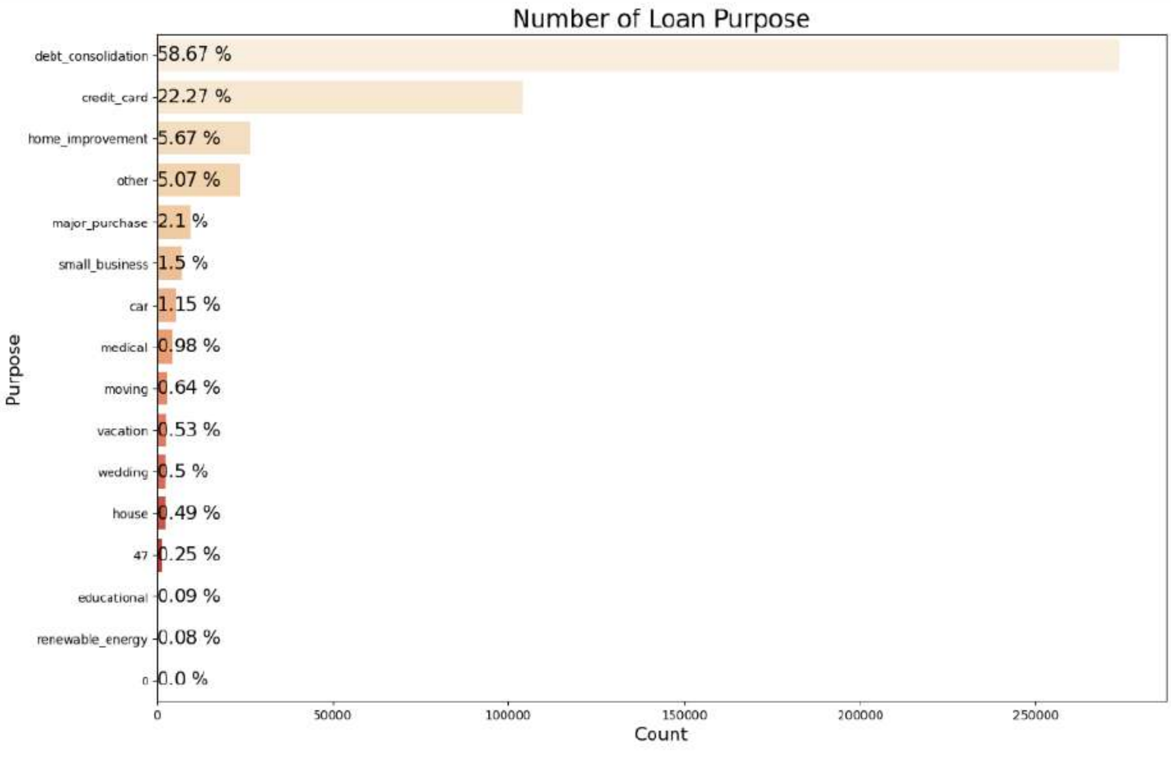
- The majority of loan status distribution is current 47.95%, and fully paid 39.54%, it means that the borrower are meeting their payment obligation.
- There are significant number of charged off 9.08%. It means that the borrower has become delinquent on payments, and potential financial loss from the lender.
- Good loan status is either current and fully paid
- Bad loan status except for these 2 things

Applicants by Borrower's Status Rate :



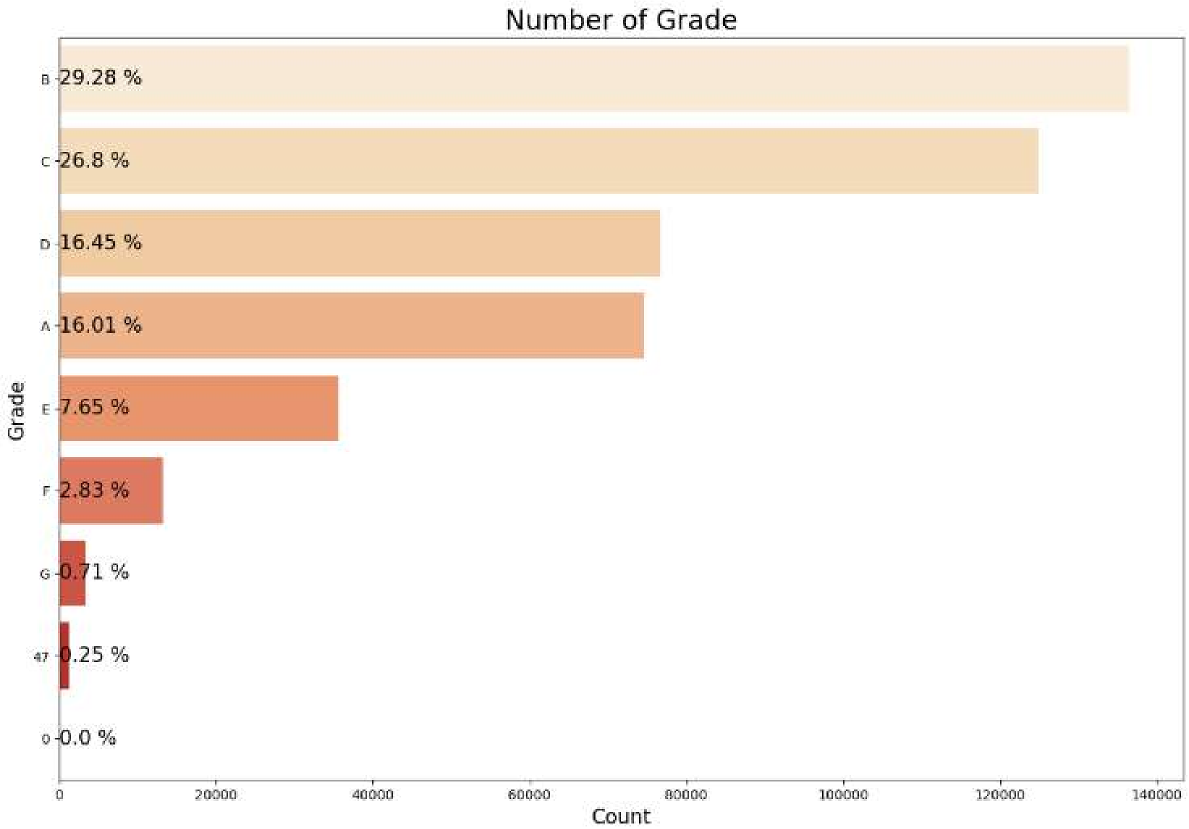
- Good loan status got high percentage with 87.49%. It means that the bank's loan performing is good.
- Bad loan status got low percentage with 12.51%. It means the bank need to analyzing the characteristic of the borrower, so they could identify early warnings sign, and implement the mitigation from failure of pay loans from customers

Applicants by Loan Purpose :



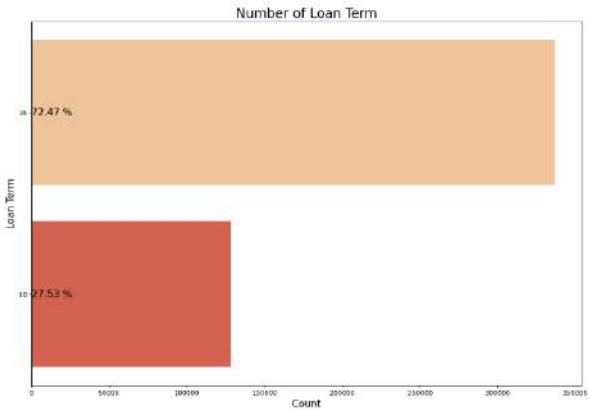
- Debt consolidation got the highest percentage for load purpose with 58.67%.
- Debt consolidation is preferred because the customer can taking out a single loan or credit card to pay off multiple debts
- The benefits of debt consolidation include a potentially lower interest rate and lower monthly payments.

Applicants by Grade :



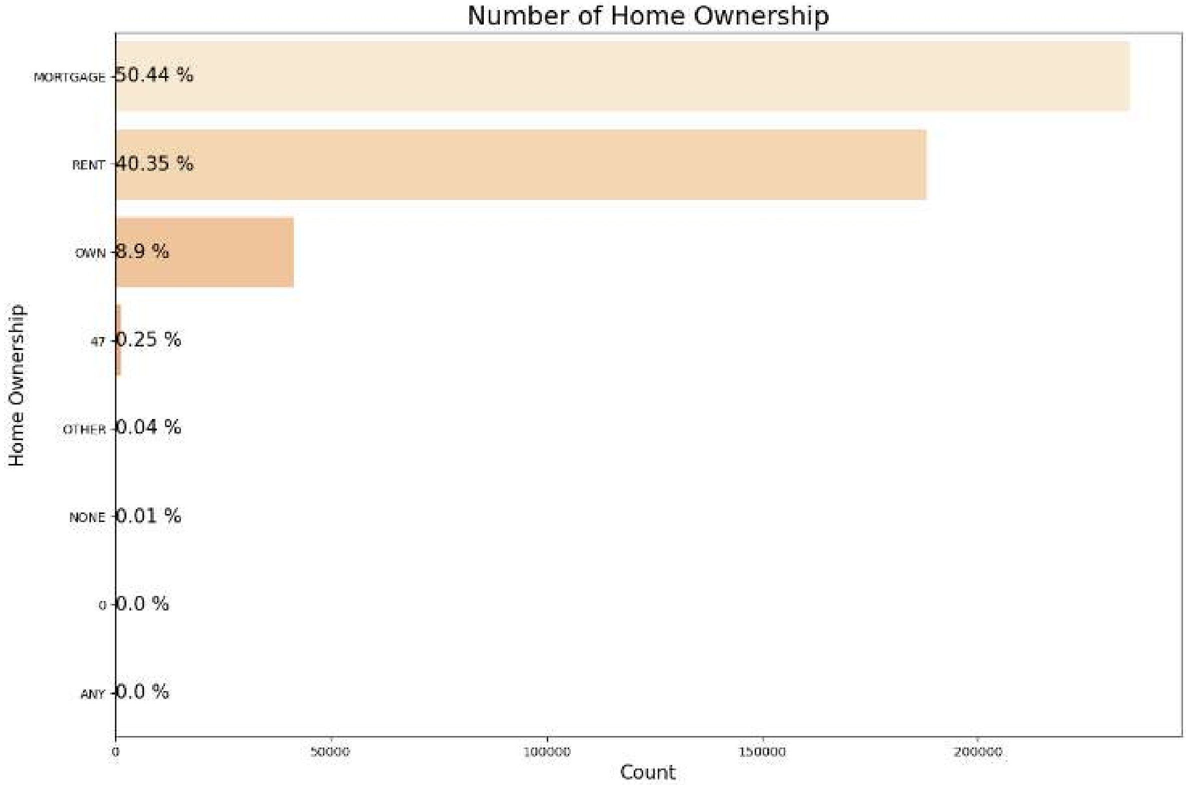
- Middle grade B and C got the highest percentage with 29.28% and 26.8%. It means that quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest are considered moderate
- Grade E,F,G got the lowest percentage. Grade E,F,G are high risk grade, because the likelihood that the borrower will repay the loan is low. So the loan company need to tighten the criteria for loan borrowers

Applicants by Loan Term :



- 36 month of loan term got the highest percentage with 72.47%. It means that short term loan are preferred by borrowers rather than long term.
- Compared to long term loans, the amount of interest paid is significantly less.
- These loans are considered less risky compared to long term loans because of a shorter maturity date.
- Short term loans are the lifesavers of smaller businesses or individuals who suffer from less than stellar credit scores

Applicants by Home Ownership :



- Mortgage got the highest percentage with 50.44%. The reason that mortgage customer is so many because a mortgage allows the customer to purchase a home without paying the full purchase price in cash.
- The second highest is rent with 40.35%. The reason that customer choose rent because no maintenance costs or repair bills, access to amenities like pool or fitness centre, no real estate taxes, and more flexibility as to where to live.
- The borrower that own their houses is only 8.9%.

Feature Engineering with Weight of Evidence (WOE) & Information Value (IV)

```
woe(df_fe_new,'initial_list_status')
```

	initial_list_status	num_observation	good_loan_prob	good_loan_prop	bad_loan_prop	weight of evidence	information_value
0	w	162846	0.899776	0.224967	0.5	-0.798654	0.340203
1	f	302258	0.864927	0.775033	0.5	0.438298	0.340203

- Weight of evidence (WOE) generally described as a measure of the separation of good and bad customers.
- Information value (IV) is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance.

Modelling

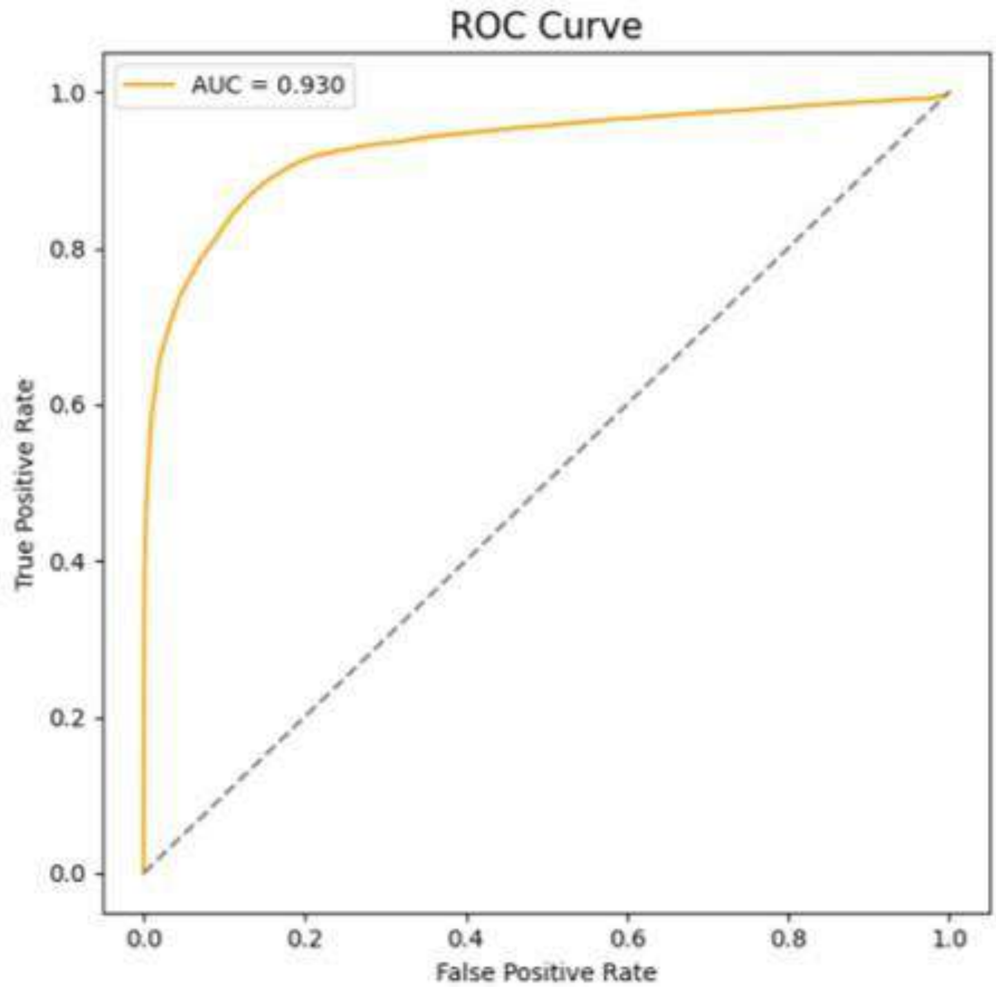
Best Parameter

```
#best parameter
search_logreg.best_params_
```

```
{'penalty': 'l2', 'C': 0.02702702702702703}
```

- Best parameter we've got is L2 (Ridge) regularization with 'C' is 0.027 which is near to 0, and leads to stronger regularization and a simpler model.

ROC/AUC



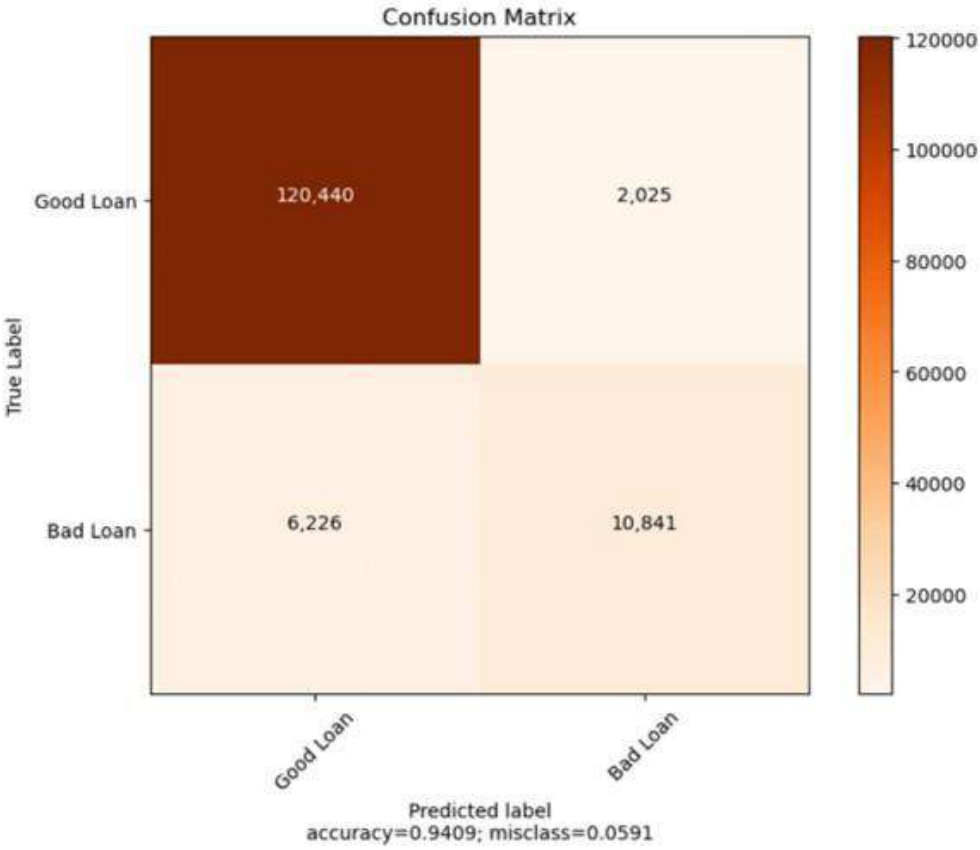
- AUC score = 0.93, which is near to 1, indicates good performance

	precision	recall	f1-score	support
0	0.95	0.98	0.97	122465
1	0.85	0.65	0.73	17067
accuracy			0.94	139532
macro avg	0.90	0.82	0.85	139532
weighted avg	0.94	0.94	0.94	139532

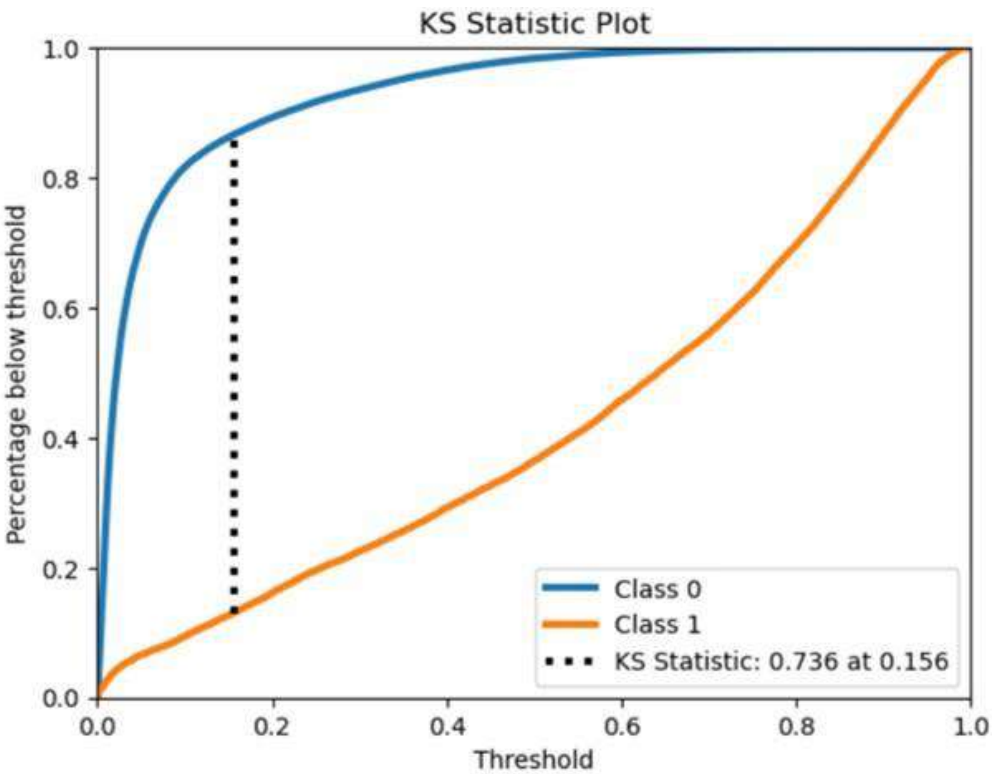
Classification Report

- Precision tells us the accuracy of positive predictions. Out of all the loan status that the model predicted would get good loan, only 85% actually did.
- Recall tells us the fraction of correctly identified positive predictions. Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for 65% of those loan status
- F1 Score = 0.73. So the model does a good job of predicting whether the loan status is considered good or bad

Confusion Matrix



- Correct classifications are the diagonal elements of the matrix 120,440 for the positive class and 10,841 for the negative class
- Accuracy rate, which is the percentage of times a classifier is correct = 94.09%



Kolmogorov-Smirnov

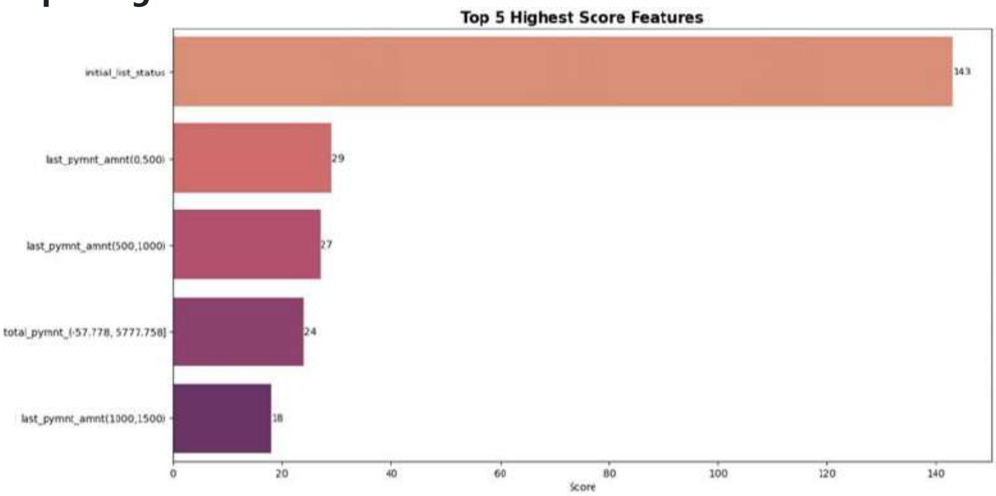
- KS Statistic = 0.736. Considered it as 'medium' dataset, which mean even though it doesn't have perfect separation, but there is enough overlap to confuse the classifier, and has wide gap between the class CDF (positive & negative instances).

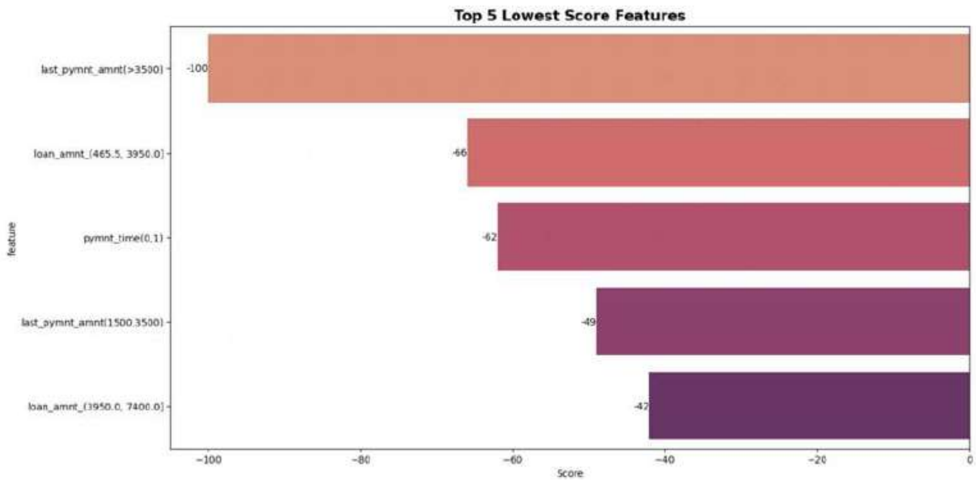
Score Card

FICO Score

- FICO score is a credit score created by the Fair Isaac Corporation (FICO)
- Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit.

Top 5 Highest & Lowest Score Features

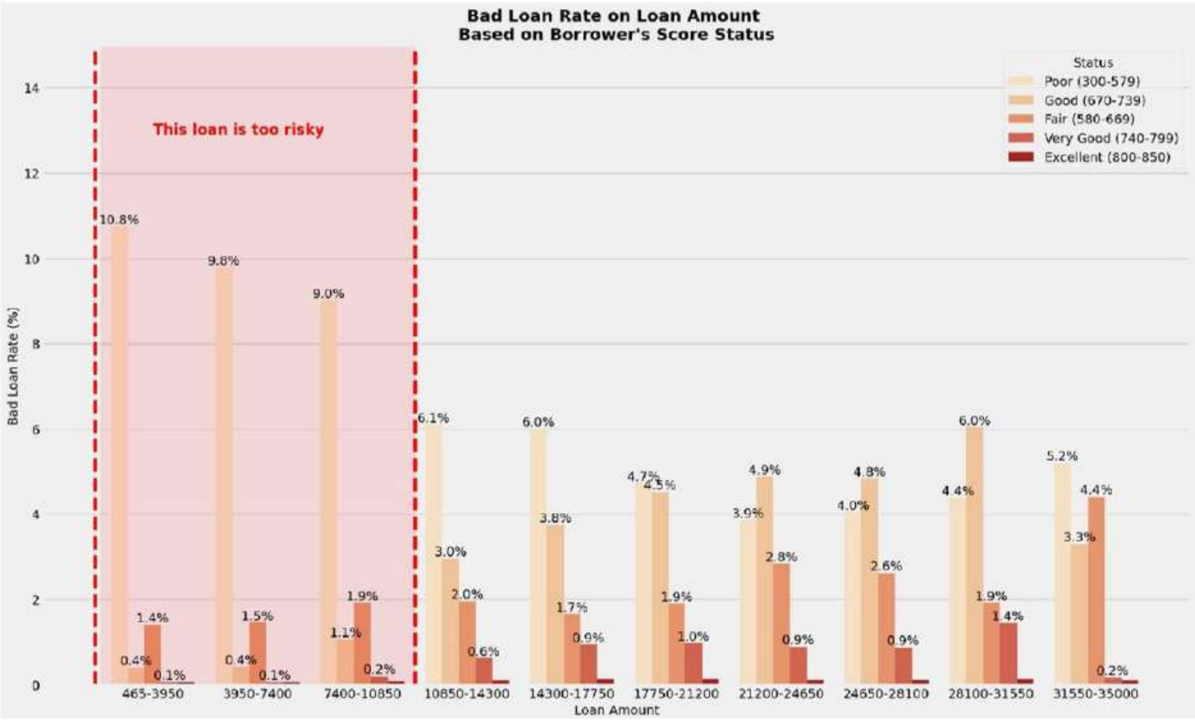




Features that make contribution to increase or decrease credit score are:

- initial list status
- last payment amount
- total payment
- loan amount
- payment time

Bad Loan Rate on Loan Amount Based On Borrower's Score Status



- Customers who have a bad credit score with a loan amount ranging from 465-10,850 have the potential to become a bad loan in the future

Recommendation

- Loan companies can build a robust and effective credit scoring model machine learning using variety of methods and criteria to assess the creditworthiness of potential customers.
- The goal is to minimize the risk of lending to individuals who are unlikely to repay their loans.
- One of method to evaluate a borrower incorporates both qualitative and quantitative measures is the 5 C's of credit (Character, Capacity, Capital, Collateral, and Conditions)

Machine Learning Project Credit Scorecard Model (Using Logistic Regression) on Home Credit Indonesia

Project Description

Problem :
The main risk for loan companies is failure to assess credit risk accurately and efficiently.
Disadvantage of Manual credit risk assessment :

- Subjectivity can introduce bias and inconsistency in decision-making
- Time-consuming especially when dealing with a large number of loan applications.
- Humans errors, such as data entry mistakes, miscalculations, or oversight of important details

Challenges :
Build a machine learning model that can automatically assess loans

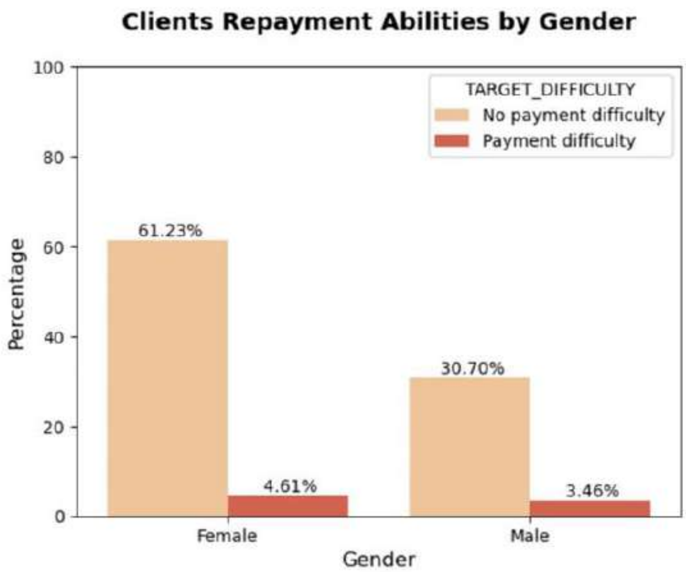
Project Result

[Click here to get full code]()

Data Preprocessing

2 Top Data Visualization & Insight

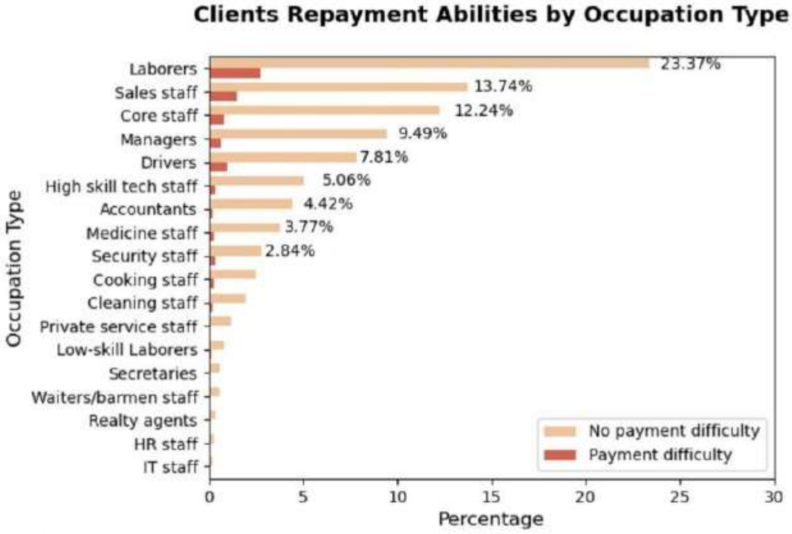
A. Clients Repayment Abilities by Gender



- 61.23% customers that do not have payment difficulty are female, and 30.70% are Male
- In UK, Women account for 65% of the home credit industry's customers (Bermeo, 2018)

- Recommendation : Start a campaign to encourage more women to apply for credit

B. Clients Repayment Abilities by Occupation Type



- 23.37% customers that do not have payment difficulty are laborers, then followed by staff and managers.
- Recommendation : Start a campaign to encourage more laborers, staff, and managers to apply for credit

Machine Learning Implementation

A. Evaluation Score

Algorithm	Mean AUROC	GINI
Decision Tree	0.5384	0.0768
Logistic Regression	**0.7304**	**0.4608**

- Mean AUROC of 0.7304 is generally considered good, indicating that the logistic regression model is effective at distinguishing between the positive and negative classes.
- Based on (Trifonova, 2012) An AUC - ROC 0.7–0.8 is considered good.
- Gini coefficient of 0.4608 indicates a relatively strong separation between the model's performance and random chance.
- It suggests that the logistic regression model has a good discriminatory ability.
- Based on (Teng, 2011) Gini coefficient 0.4 - 0.5 considered big gap.

B. Score Card

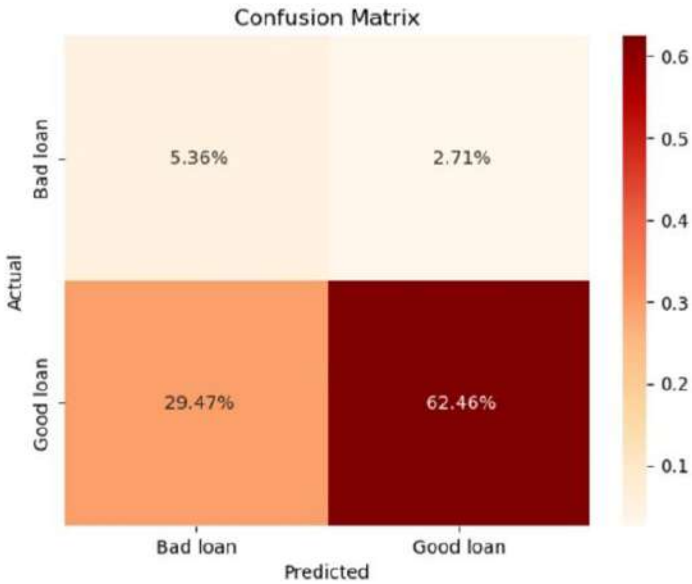
Ori_Feature_Name	Score_Calculation
intercept	555.0
CODE_GENDER	-8.0
CODE_GENDER	11.0
NAME_EDUCATION_TYPE	60.0

- Base (Intercept) = 555
- Min Score = 300 (FICO)
- Max Score = 850 (FICO)

C. Confusion Matrix with Threshold = 0.5

	precision	recall	f1-score	support
0	0.15	0.66	0.25	4965
1	0.96	0.68	0.80	56538
accuracy			0.68	61503

- Precision = Out of all the loan status that the model predicted would get good loan, only 96% actually did.
- Recall = Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for 68% ofthose loan status.
- F1 Score = 0.8. F1 score of 0.7 or higher is often considered good (spotintelligence.com, 2023) The accuracy is not really good because we've got 0.68 out of 1



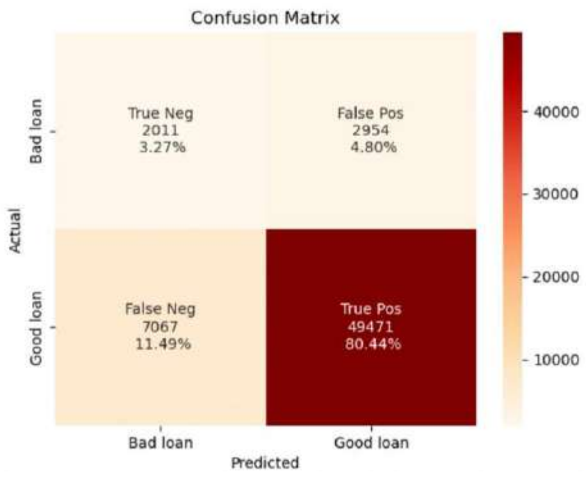
62.46% got correct variable of good loan

D. Confusion Matrix with Best Threshold

- Best threshold = 0.353918 (Using Youden J-Statistic)
- Best threshold is used to minimized the False Positive Rate and maximize the True Positive Rate

	precision	recall	f1-score	support
0	0.22	0.41	0.29	4965
1	0.94	0.88	0.91	56538
accuracy			0.84	61503

- Precision = Out of all the loan status that the model predicted would get good loan, only 94% actually did.
- Recall = Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for 88% ofthose loan status.
- F1 Score = 0.9. F1 score of 0.7 or higher is often considered good (spotintelligence.com, 2023) The accuracy increased significantly from 0.68 to 0.84



80.44% got correct variable of good loan.

E. Approval & Rejection Rate

(Threshold = 0.5)

threshold	score	n_approved	n_rejected	approval_rate	rejection_rate
0.500012	551.0	40082	21421	0.651708	0.348292

- Choosing a 0.5 threshold might mean rejecting a lot of applicants with rejection rate 34%, which could lead to losing business

(Best Threshold = 0.353918)

threshold	score	n_approved	n_rejected	approval_rate	rejection_rate
0.353918	516.0	52426	9077	0.852414	0.147586

- With best threshold, we've got rejection rate 14%
- So, we've decided to keep our preferred threshold = 0.353918 and Credit Score of 516

Business Recommendation

Partial Auto Reject & Auto Approve

- If a submission seems bad, it is rejected right away.
- If a submission appears to be very good, it is accepted immediately.
- If there's uncertainty, it is manually checked by the assessment team.

Create targeted campaign

- We should launch additional campaigns targeting women, laborers, staff, and managers to encourage them to apply for credit.

Machine Learning Project Using Logistic Regression to Predict Absenteeism

Project Description

Problem :
Absenteeism of workers refers to the habitual pattern of absence from work, often without any valid reason. It can be a significant challenge for organizations as it can disrupt workflow, reduce productivity, and impact overall team morale. Understanding the reasons behind absenteeism is crucial for developing effective strategies to address and mitigate this issue.

Challenges :
Build a machine learning model that can predict the causes of absenteeism

Project Result

[Click here to get full code]()

A. Logistic Regression

- Logistic Regression is type of classification
- We create 2 class ==> Moderately absent, Excessively absent
- We'll take the median value of Absenteeism Time in Hours, and use it as cut-off line
- Absenteeism Time in Hours < Median ==> Moderately absent
- Absenteeism Time in Hours > Median ==> Excessively absent

```
#check if dataset is balanced (what % of targets are 1s)
(data_preprocessed['Excessive Absenteeism'].sum() / data_preprocessed.shape[0])*100
```

```
46.09826589595375
```

- There are 46% of target 1 (Absenteeism Time > 3 hours)
- A balance of 45-55 is almost always sufficient

Standardize the data
Standardize numerical data using
StandardScaler

Reason_1	Reason_2	Reason_3	Reason_4	Day of the Week	Transportation Expense	Age	Daily Work Load Average	Children	Pets	Day of the Week_std	Transportation Expense_std	Age_std	Daily Work Load Average_std
0	0	0	1	1	289	33	239.554	2	1	-0.685271	0.998885	-0.529189	-0.803696
0	0	0	0	1	118	50	239.554	1	0	-0.685271	-1.582804	2.126691	-0.803696

Logistic Regression

Feature name	Coefficient	Odds_ratio
Reason_3	3.018118	20.452765
Reason_1	2.889772	17.989206
Reason_4	1.026263	2.790617
Reason_2	0.826736	2.285846
Transportation Expense_std	0.661453	1.937605
Children_std	0.348756	1.417303
Daily Work Load Average_std	0.238527	1.269378
Education	0.133734	1.143089
Month Value_std	0.095475	1.100181
Distance to Work_std	-0.029027	0.971391
Body Mass Index_std	-0.029027	0.971391
Age_std	-0.227760	0.796315
Day of the Week_std	-0.251436	0.777683
Pets_std	-0.352381	0.703012
Intercept	-1.731263	0.177061

A feature is not particularly important :

- if its coefficient is around 0 & its odds ratio is around 1
- A weight (coefficient) of 0 implies that no matter the feature value, we will multiply it by 0 (in the model)
- For a unit change in the standardized feature, the odds increase by a multiple equal to the odds ratio (1 = no change)
- Example odds x odds_ratio = new_odds ==> 5:1 x 1 = 5:1 (no change)

The variable that has its coefficient is around 0 & its odds ratio is around 1 ==> USELESS for our model:

- Education
- Month Value
- Distance to Work
- Body Mass Index

  Pet is at the bottom of the table, but their weights are still far away from 0, it's indeed important.
  Pet is continous variable, that has negative coefficient (-0.352381). Its odds is (1 - 0.703012)*100 = 29.6% lower than the base model(no pet)

Interpreting the coefficient :
The further away from 0 a coefficient is, the bigger its importance The highest odds_ratio that affect Absenteeism are :

- Reason 3 (poisoning)

- Reason 1 (diseases)
- Reason 4 (light reasons for absence)
- Reason 2 (pregnancy)

So, the most crucial for excessive absence is positioning. The odds of someone being excessively absent after being poisoned is 20 times higher than when no reason was reported (Reason 0 ==> Baseline model)

Accuracy of the model:
After drop weak variable :

```
#train accuracy of the model
reg2.score(x_train2, y_train2)
```

```
0.7540687160940326
```

Based on the data we used, our model learned to classify 75.40% of the observation correctly

Machine Learning Project Using KMeans to Country Clustering

Project Description

Problem :
In our increasingly interconnected world, understanding global patterns based on geographical location and linguistic diversity is crucial for various applications. This project aims to leverage the geographic coordinates (latitude and longitude) and primary language spoken in each country to group them into clusters.

Challenges :
Build a machine learning model that can clustering country based on their regional & language.

Project Result

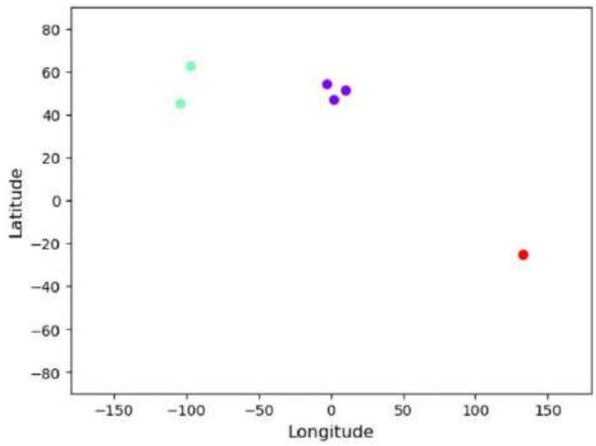
Click here to get full code

Simple K-Means Clustering

Dataset :

Country	Latitude	Longitude	Language
USA	44.97	-103.77	English
Canada	62.40	-96.80	English
France	46.75	2.40	French
UK	54.01	-2.53	English
Germany	51.15	10.40	German
Australia	-25.45	133.11	English

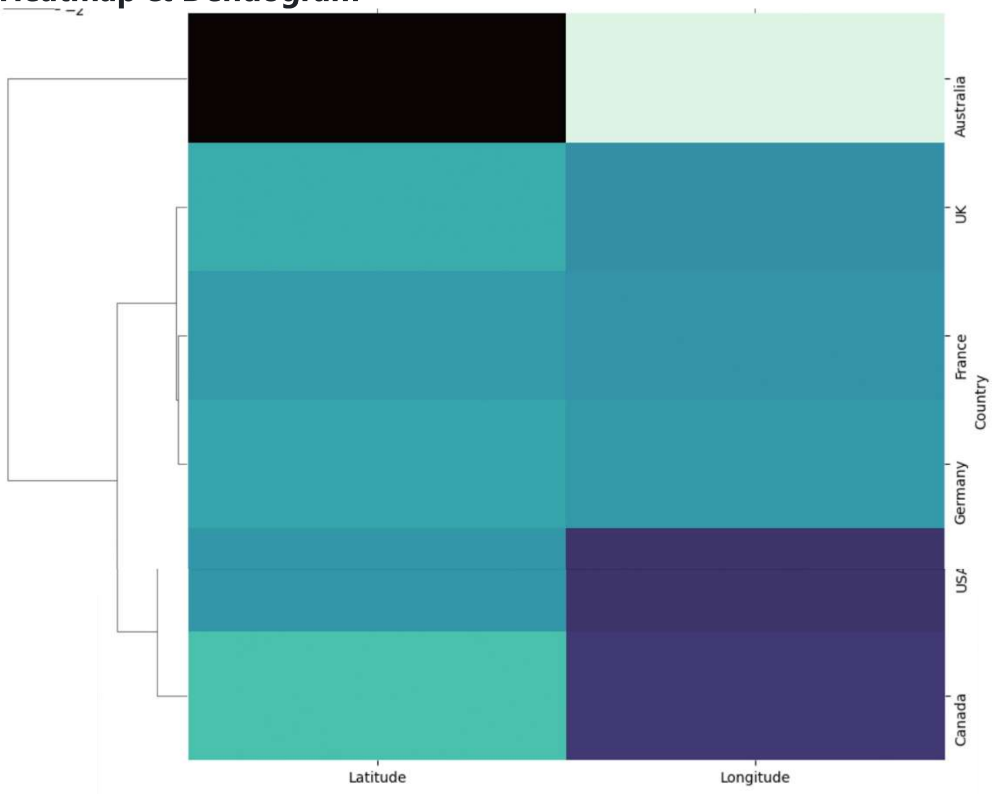
A. Clustering Based On Longitude and Latitude



From 6 country, we clustering them based on their geographical location, so we select the features latitude and longitude. There are 3 clusters

- green dots = USA & Canada
- purple dots = France, UK, Germany
- red dots = Australia

Heatmap & Dendogram



There are 2 features :

- Latitude
- Longitude

There are 6 observations :

- Australia
- UK
- France
- Germany
- USA
- Canada

Insight :

- In terms of Latitude, only Australia has different color (dark blue), because other country is located in Northern Hemisphere, while Australia is located in Southern Hemisphere
- Germany, France, and UK has similar color ==> cluster 1
- USA, and Canada has similar color in terms of longitude, and slight difference on latitude ==> cluster 2

- Australia is compeletely different in both latitude and longitude => cluster 3

*B. Clustering Based On Language:**

Clustering is about :

- minimizing the distance between points in a cluster
- maximizing the distance between clusters distance between points in a cluster ==> WCSS (Within Cluster Sum of Squares)

If we minimize WCSS, we have reached the perfect clustering solution But there's problem:

- observation: 1,000, cluster = 1,000, WCSS=0 (min)
- observation: 1,000, cluster = 1 WCSS= max
- What we want : WCSS middle Ground ==> - observation: N, cluster = small, WCSS= low ==> Elbow Method
- **Elbow Method** ==> the biggest number of clusters for which still getting significant decrease in WCSS

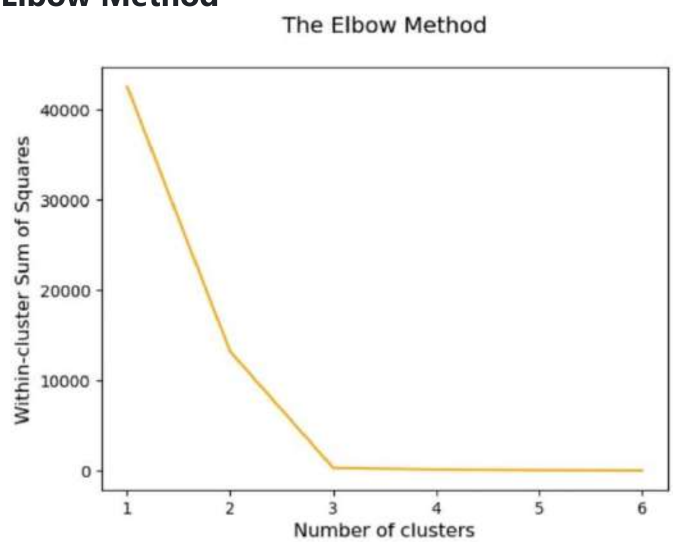
WCSS (Within-Cluster Sum of Squares)

```
wcss
```

```
[42605.41356666667,
 13208.958119999996,
 290.10523333333333,
 113.91233333333332,
 39.00624999999998,
 0.0]
```

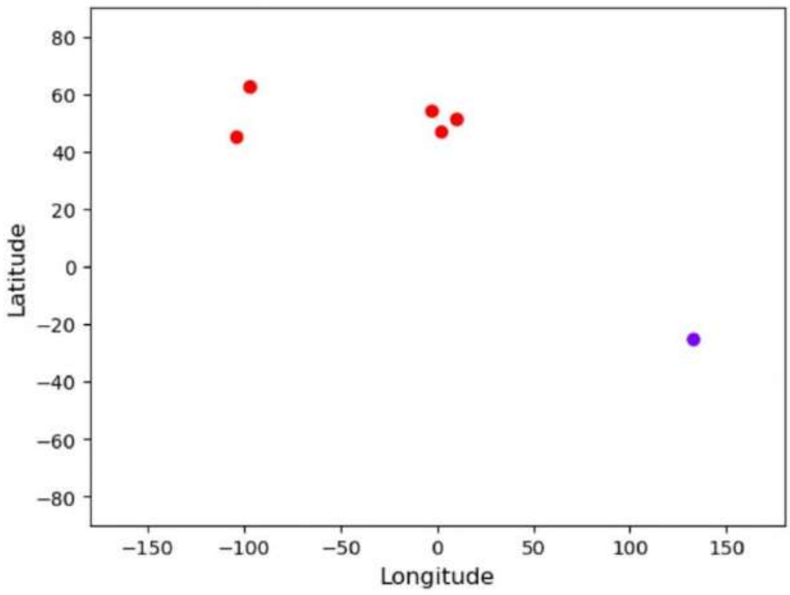
We've got 6 WCSS and then we're gonna implement it to The Elbow Method to find the best of number of clustering to use.

Elbow Method



- 3 cluster is the best solution
- 2 cluster solution would be suboptimal as the leap from 2 to 3 is very big in terms of WCSS.

In this project, we plot 2 cluster based on their language :



Machine Learning Project Using Linear & Logistic Regression to Predict GPA from SAT Scores

Project Description

Problem :
Understanding the relationship between standardized test scores and academic performance is essential for educational institutions to make informed admission decisions. By leveraging historical data, the goal is to create a tool that assists admission offices in evaluating the potential academic success of applicants and provides valuable insights into the predictive power of SAT scores.

Challenges :
Build a machine learning model that can predict GPA from SAT Scores

Project Result

[Click here to get full code]

Dataset

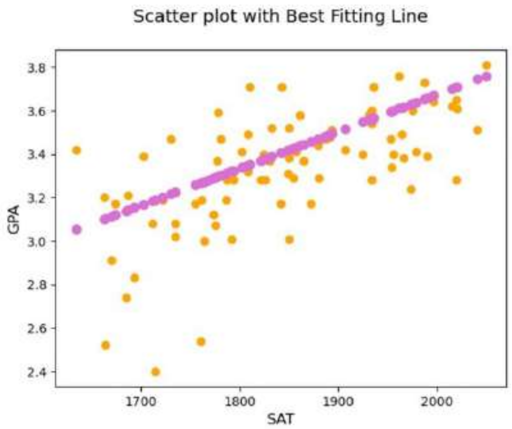
	SAT	GPA	Rand 1,2,3	Attendance	Admitted	Gender
0	1714	2.40	1	No	No	Male
1	1664	2.52	3	No	Yes	Female

- There are 84 students who have studied in college
- SAT Score = Critical reading + Mathematics + Writing

- GPA = Grade Point Average (at graduation from university

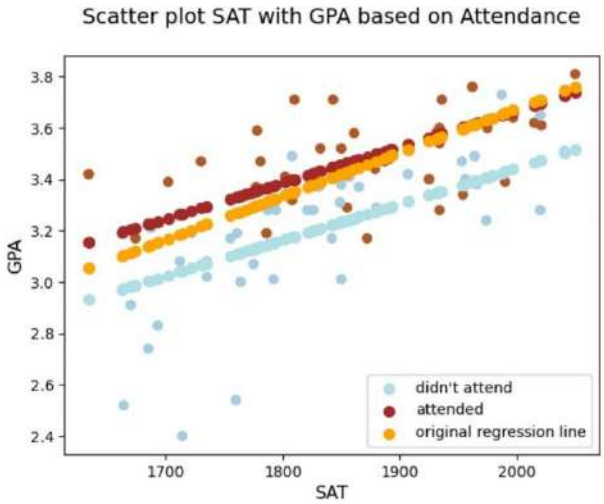
Linear Regression

GPA based on SAT Score



- That is the best fitting line, or the line which is closest to all observation simultaneously
- Example if there is student who has SAT score 1700, then he will got GPA 3.165
- There is strong relationship between SAT and GPA
- The higher the SAT of a student, the higher their GPA

GPA based on SAT & Attendance



- From this dataset, we found that average of students attendance more than 75% of lectures is only 46.42% have attended. Mean < 0.5 shows that there are more 0s than 1s.
- On average the GPA of those who attendeded is higher than the one didn't attend the class.

Making Predictions

Prediction 1

Create prediction of 2 students, whose the one that get higher GPA :

- Budi, who got 1700 on SAT and did not attend >75% of lecturers
- Ani, who got 1670 on SAT and attended >75% of lecturers

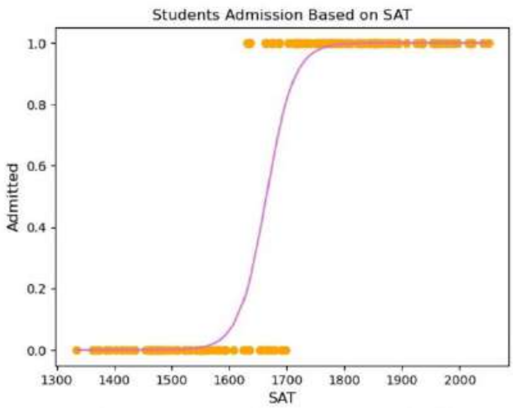
	const	SAT	Attendance	Predictions
Budi	1	1700	0	3.023513
Ani	1	1670	1	3.204163

- The predicted GPA at graduation for Budi is 3.02
- The predicted GPA at graduation for Ani is 3.20

- Ani scored lower on SAT, but she attended > 75% of lectures, and she is predicted to graduate with a significantly higher GPA than Budi.

Logistic Regression

Predicting whether student will be admitted or not



- This function shows the probability of admission given an SAT score
- When SAT score is relatively low, the probability of getting admitted is 0%
- When SAT score is relatively high, the probability of getting admitted is 100%
- Score between 1,600 and 1,750 is uncertain
- SAT score 1,650, the students roughly 50% chance of getting in
- SAT score 1,700, the students got 80% chance of getting in

Predicting which gender will be the most admitted

Logit Regression Results

Dep. Variable:	Admitted	**No. Observations:**	168
Model:	Logit	**Df Residuals:**	165
Method:	MLE	**Df Model:**	2
Date:	Wed, 13 Dec 2023	**Pseudo R-squ.:**	0.8249
Time:	20:39:48	**Log-Likelihood:**	-20.180
converged:	True	**LL-Null:**	-115.26
Covariance Type:	nonrobust	**LLR p-value:**	5.118e-42

	coef	std err	z	P>|z|	[0.025	0.975]
const	-68.3489	16.454	-4.154	0.000	-100.598	-36.100
SAT	0.0406	0.010	4.129	0.000	0.021	0.060
Gender	1.9449	0.846	2.299	0.022	0.287	3.603

```
#take the exponential of gender
np.exp(1.9449)
```

```
6.992932526814459
```

- odds of female to get admitted are 6.99 times odds of male
- given the same SAT score, a female has 7 times higher odds to get admitted than the male
- in this particular university (degree), it is much easier for females to enter
- example communications, most of them are female, while STEM predominantly male

Accuracy

```
#calculate accuracy of the model
accuracy_train = (cm[0,0] + cm[1,1]) / cm.sum()
accuracy_train
```

```
0.9464285714285714
```

- The accuracy of our model is 94.64%. Our model seems good at classifying

Machine Learning Project Using Linear Regression to Predict Price of Used Car

Project Description

Problem :
Buying or selling a used car can be a complex process, and determining a fair market value for a used car is often subjective. This project addresses the challenge of predicting the price of a used car based on its specifications.

Challenges :
Build a machine learning model that can predict the price of used car

Project Goal

By leveraging machine learning techniques and historical data, the goal is to develop a model that provides accurate and reliable estimates of a used car's market value, taking into account various features and attributes.

Project Result

[Click here to get full code]

Dataset

Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year	Model
BMW	4200.0	sedan	277	2.0	Petrol	yes	1991	320
Mercedes-Benz	7900.0	van	427	2.9	Diesel	yes	1999	Sprinter 212

- Brand ==> BMW is generally more expensive than Toyota
- Mileage ==> the more car is driven, the cheaper it should be
- EngineV ==> sports car have larger engines than economy cars
- Year of production ==> the older the car, the cheaper it is, with exception of vintage vehicles

Check linearity

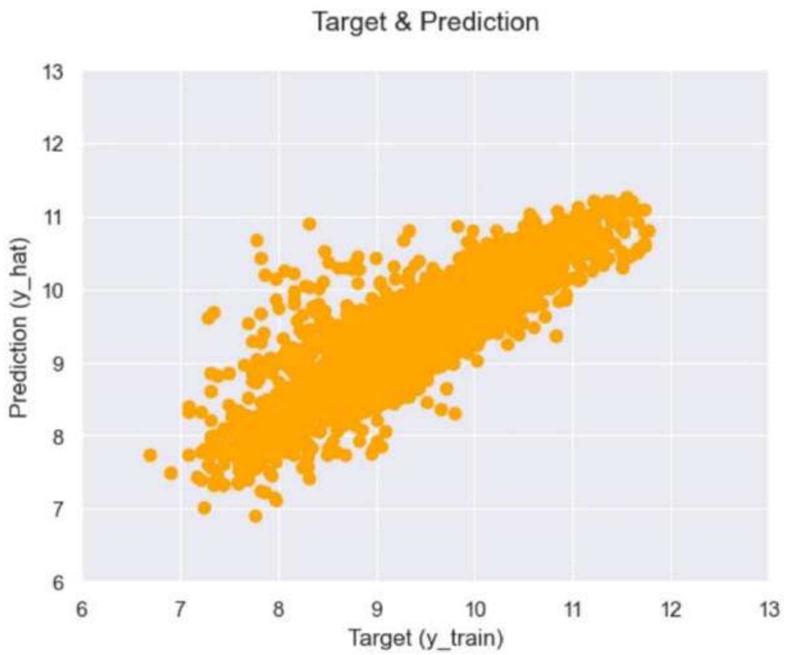
- OLS assumption of regression is linear, but from the plot, price is exponentially distributed
- good transformation in that case is a log transformation linear

Check Multicollienarity

VIF	Features
3.791584	Mileage
10.354854	Year
7.662068	EngineV

- Year has the highest VIF, drop 'Year'
- VIF 'Year' = 10.35 > 10 ==> unacceptable

Linear Regression Model



- For high prices, we have a higher concentration of values around the 45 degree line ==> Our model is very good at predicting higher prices.

```python
#R-squared of the model
reg.score(x_train, y_train)
```

```
0.744996578792662
```

- Our model is explaining 75% of the variablity of the data, relatively good result

Features	Weights
Mileage	-0.448713
EngineV	0.209035
Brand_BMW	0.014250
Brand_Mercedes-Benz	0.012882
Brand_Mitsubishi	-0.140552
Brand_Renault	-0.179909
Brand_Toyota	-0.060550
Brand_Volkswagen	-0.089924
Body_hatch	-0.145469
Body_other	-0.101444
Body_sedan	-0.200630
Body_vagon	-0.129887
Body_van	-0.168597
Engine Type_Gas	-0.121490
Engine Type_Other	-0.033368
Engine Type_Petrol	-0.146909
Registration_yes	0.320473

Weights Interpretation :

- A positive weight shows that as a feature increases in value, so do log_price and 'Price' respectively
- Example : EngineV, the bigger the Engine volume, the higher the price
- A negative weight shows that as a feature increases in value, log_price and 'Price' decrease
- Example : Mileage, the more a car is being driven, the lower the price gets
- Dummies are compared with the benchamrk dummy
- A positive weight shows that the respective category is more expensive than the benchmark
- Example : respective category (BMV) is more expensive than the benchmark (Audi)
- A negative weight shows that the respective category is less expensive than the benchmark
- Example : respective category (Mitsubishi) is more expensive than the benchmark (Audi)
- The bigger the weights the bigger the impact
- Mileage is te most prominent feature in the regression. It is more than twice as important as Engine Volume

Machine Learning Project Build a Movie Recommendation System

Project Description

Problem :
The Movie Recommendation System project aims to develop an intelligent system that suggests personalized movie recommendations to users based on their preferences and viewing history.

Challenges :
Build a machine learning model that can recommend movie based on user preference.

Project Result

[Click here to get full code]

3 Types of Recommendation System :

1. Popularity based recommendation system
   Recommend list of popular movie.
   To get list of popular movie in this dataset, we calculate weighted rating, and here is the result :

   Top 10 Popular Movie:

	title	weighted_rating
1881	The Shawshank Redemption	15636.015145
3337	The Godfather	15452.408618
2731	The Godfather: Part II	15269.183637
2294	Spirited Away	15268.992360
3865	Whiplash	15268.858331
1818	Schindler's List	15268.835976
3232	Pulp Fiction	15268.110923
662	Fight Club	15268.015418
2247	Princess Mononoke	15086.010823
1987	Howl's Moving Castle	15086.004701

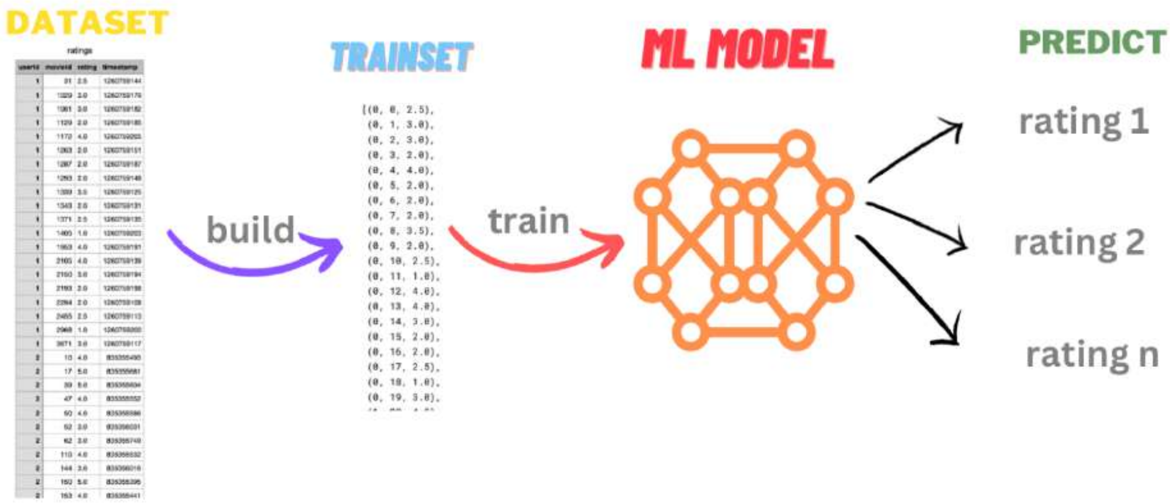
2. Content based filtering
   When click certain movie, it will give recommendation of similar movie.
   To get list of similar movie, we use Term Frequency & Inverse Document Frequency.

For example, we try to know 3 movies that is similar with movie title "John Carter" :

3 Top movie title that is similar with "John Carter" is :

- 'Get Carter'
- 'The Marine 4: Moving Target'
- 'Raising Cain'

3. Collaborative filtering
   Predict what rating the user gonna give.



Example :

What rating of user 15 will give to movie id 1956?

```
svd.predict(15, 1956).est
```

```
3.4391919289891364
```

- The user with id 15 predicted will give ratings 3.49 to movie id 1956
- The ratings quite good because the rating ranges from 1 to 5.

Machine Learning Project of House Price Prediction

Project Description

The House Price Prediction project involves the development of a machine learning model to predict residential property prices based on various features. Utilizing a dataset containing information such as square footage, number of bedrooms, location details, and other relevant attributes, the project aims to build an accurate and reliable predictive model.

Challenges :
Build a machine learning model that can predict House Price Prediction.

Project Result

[Click here to get full code]

Dataset Used

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178
538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1991	98125
180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028

Even though in this dataset, there is categorical values, for example waterfront, and view columns, but it's already in numeric version (1, 0). So we do not need to do label encoding, just standardization using StandardScaler.

Modelling

I already tried several algorithm such as :

- Linear Regression
- Hyperparameter Tuning using Ridge, Lasso, and ElasticNet
- Decision Tree
- Random Forest
- Support Vector Regressor

But The **Best Evaluation Score** is using **Linear Regression model**.

```
eval_regression(regressor)
```

```
RMSE(test):  208296.7277211889
RMSE(train):  198133.94425362692
MAE(test):  127486.80255718411
MAE(train):  124691.93980379181
MAPE(test): 0.253044843464385
MAPE(train): 0.2544052061162715
r2(test): 0.6994627057969898
r2(train): 0.6995155846436756
r2_cross_val_test:  0.6945908283283323
r2_cross_val_train: 0.7002121455769499
```

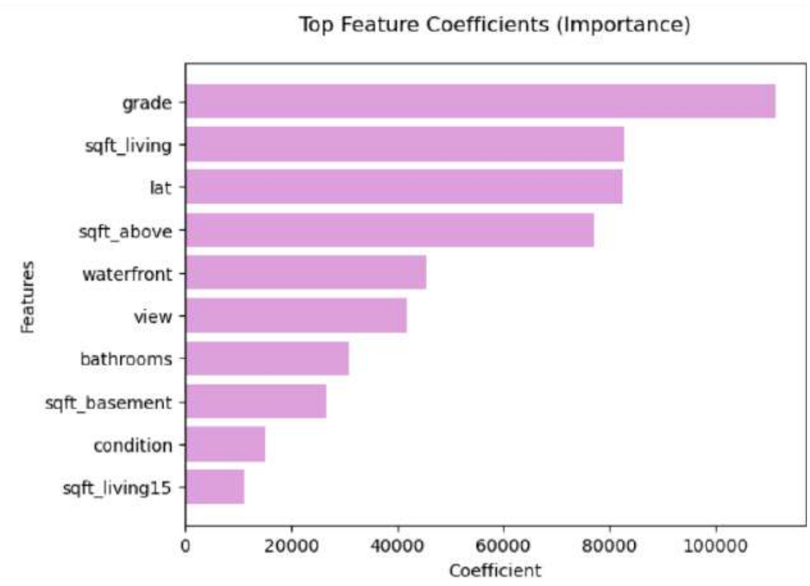
- The r2 score between the train data and test data is not far, so the model does not overfit.
- The error from this model is small 25%.

House Price Prediction using Best Model

From the best model, we could predict the House Price using Linear Regression :

	Actual Price	Prediction Price
0	365000.00	458597.07
1	865000.00	748993.76
2	1038000.00	1243303.76
3	1490000.00	1665116.95
4	711000.00	737302.06
5	211000.00	283239.59
6	790000.00	831732.88
7	680000.00	495383.02
8	384500.00	385779.82
9	605000.00	474179.42

Feature Importance of House Price Prediction



Top 10 of features that have the most significant impact on the predicted target variable :

1. 'grade',
2. 'sqft_living',
3. 'lat',
4. 'sqft_above',
5. 'waterfront',
6. 'view',
7. 'bathrooms',
8. 'sqft_basement',
9. 'condition',
10. 'sqft_living15',

Uber New York Data Analysis

Project Description

Data analyst got a project to analyze Uber pickups on New York to get insight from this data. The challenges of these project include :

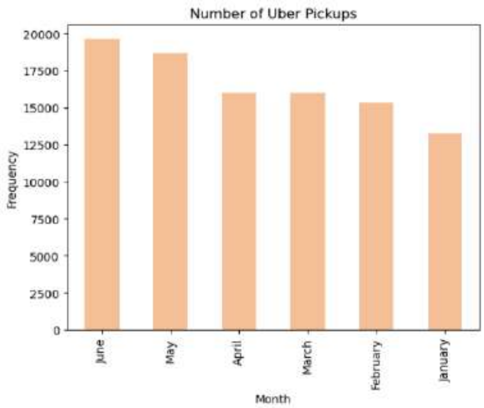
- Analyzing which month have max uber pickup
- Analyzing Hourly Rush in New York
- Analyzing Most Active Uber Base Number
- Perform Spatial Analysis to find what locations of New York City are getting Rush
- Perform Pair Wise Analysis to Examine Rush Hour

Project Result

[Click here to get on full code]

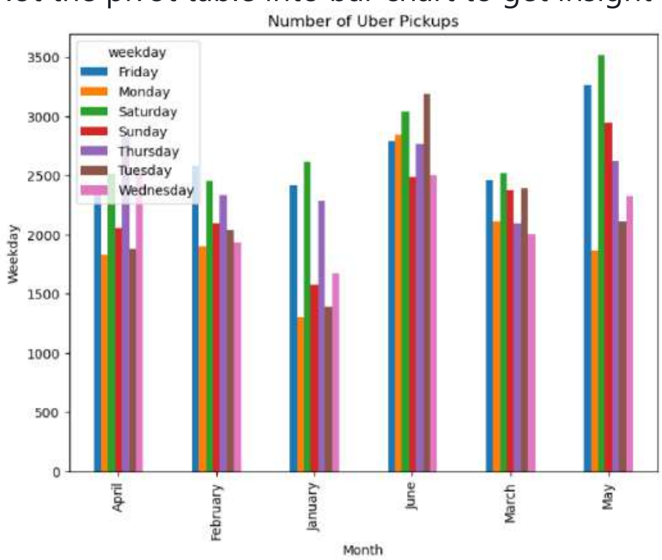
Analyzing which month have max uber pickup

Create new dataset that includes Month, and Frequency of Uber pickups. Then plot the dataset into the bar chart.



Insight :
June seems to have max number of pickups on Uber

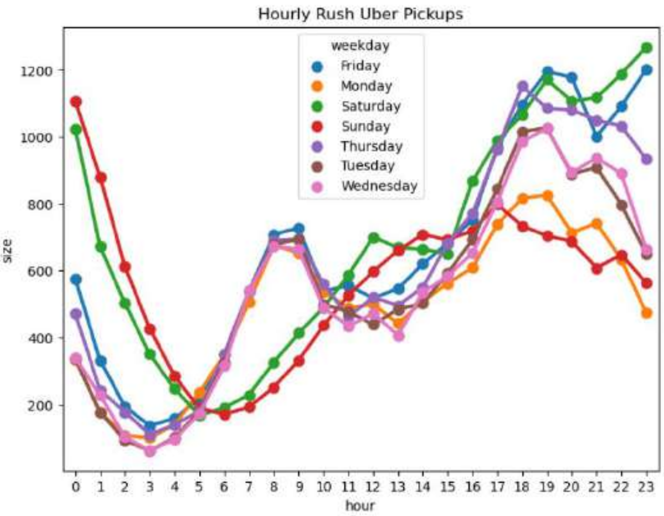
Plot the pivot table into bar chart to get insight what is the highest weekdays of uber pickups



Insight :

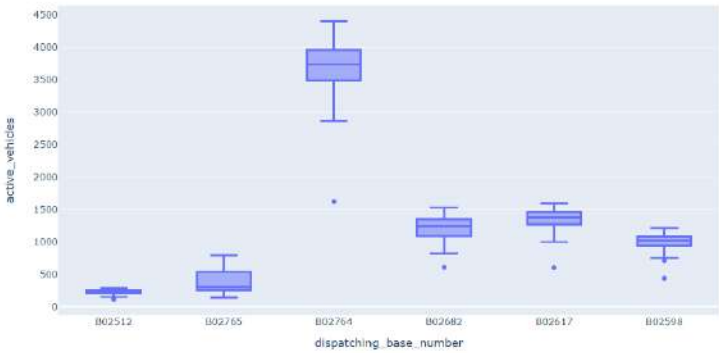
- The highest number of pickups in each month is on Friday and Saturday.
- It seems people on New York hanging out a lot on these days, like go shopping, spend time with family outside, etc.

Analyzing Hourly Rush in New York



Insight :
It seems Saturday and Sunday has similar trend in late night, morning, and afternoon. But in the evening (starts from 17:00) they exhibits opposite trend, where Saturday pickup continue to increasing, but Sunday pickupstakes a downward turn.

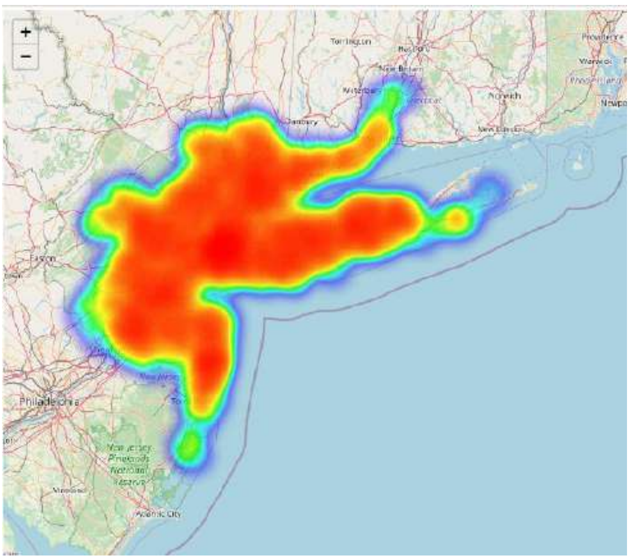
Analyzing Most Active Uber Base Number



Insight :
The most active Uber base number is B02764 with 5 summary stats of data :

- min = 2,862 active vehicles
- q1 (25th percentile) = 3,483 active vehicles
- q2 (50th percentile) = 3,734 active vehicles
- q3 (75th percentile) = 3,957 active vehicles
- max = 4,396 active vehicles

Perform Spatial Analysis to find what locations of New York City are getting Rush



Insight :
Midtown Manhattan is the locations of New York City that are getting Rush, because from this heatmap, it is clearly huge bright spot.
The reason maybe because Manhattan is the most densely populated of New York City’s.

Bitcoin Data Analysis

Project Description

Data analyst got a project to analyze price of Bitcoin to get insight from this data. The challenges of these project include :

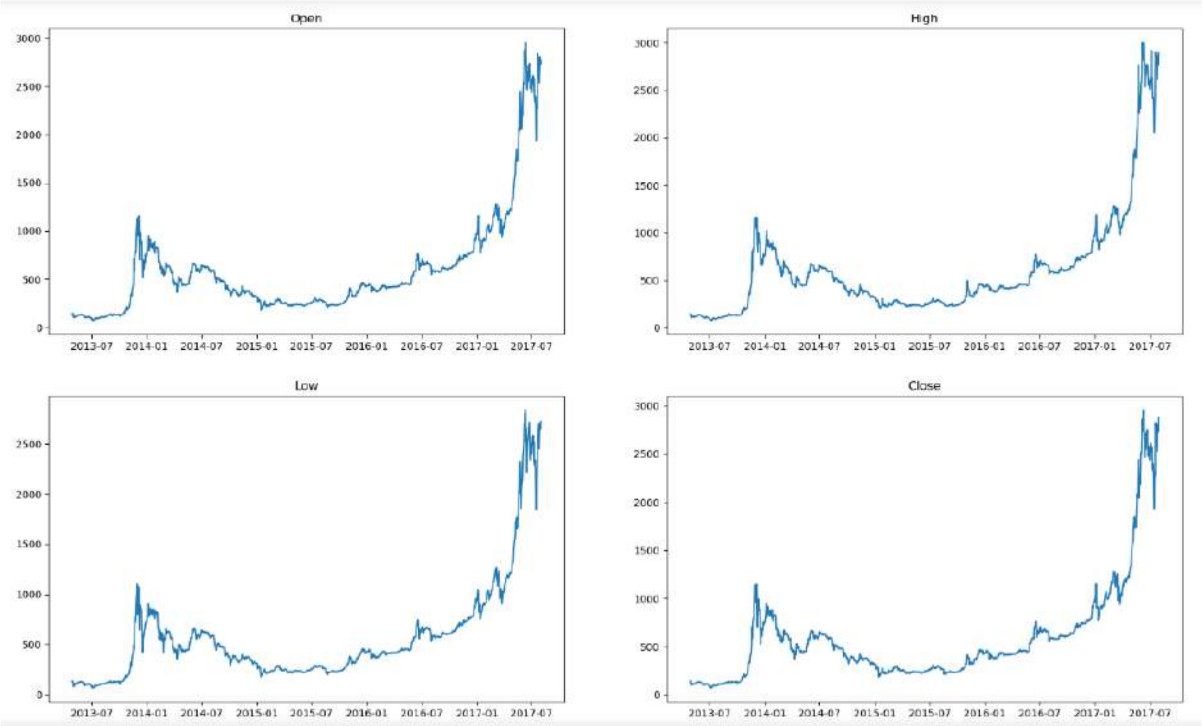
- Analyzing Change in Price of The Bitcoin Overtime
- Analyzing Bitcoin Price Using Candle-Stick Chart
- Analyzing Closing Price in depth
- Perform Analysis on Closing Price on Yearly, Quarterly, and Monthly Basis
- Analyzing Daily Change In Closing Price of Bitcoin

Project Result

Click here to get on full code

Analyzing Change in Price of The Bitcoin Overtime

Then plot Open, High, Low, Close price into the line chart.



Insight :
For each price for Open, High, Low, and Close, there are spike in 2014 and 2017.

Analyzing Bitcoin Price Using Candle-Stick Chart

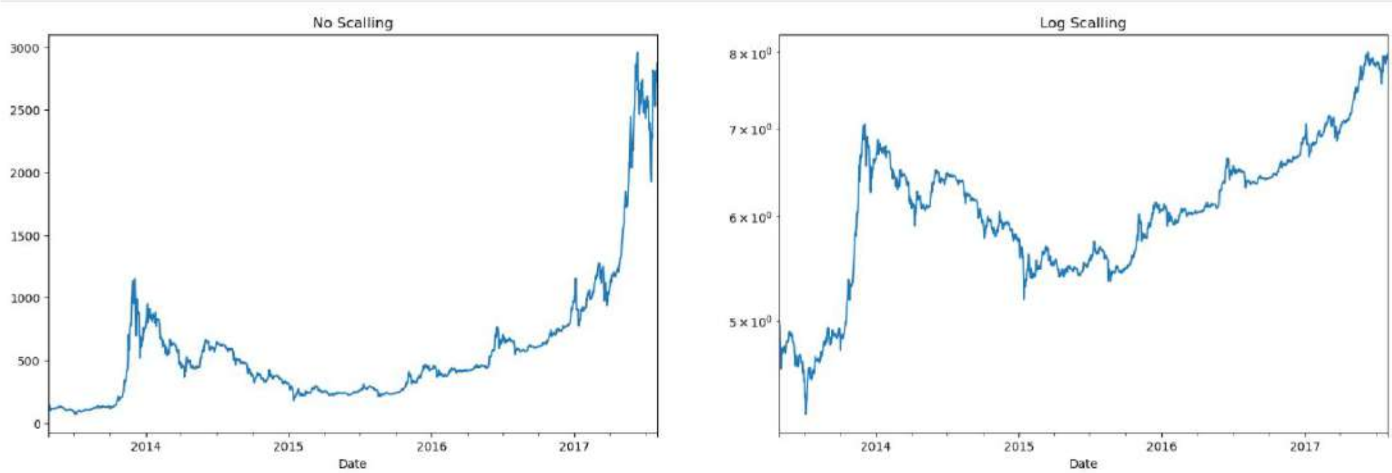
Create sample data from bitcoin dataset. Then create candle stick, where the x-axis is 'Date' and price data = 'High', 'Open', 'Close', 'Low'.

Insight :
On 1 May 2013, we got the result that:

- open price : 139
- high price : 139.89
- low price : 107.72
- close price : 116.99

Analyzing Closing Price in depth

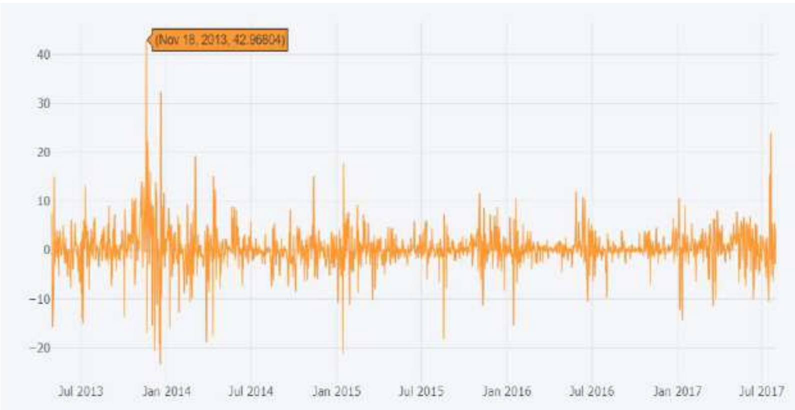
Create the linear scale and log scale for close price



Insight :

- Logarithmic price scale are better than linear scale for showing less severe price increases or decreases.
- There is an upward trend in 2014 and 2017 for each graph.

Analyzing Daily Change In Closing Price of Bitcoin



Insight :

- On 18 November 2013 there is spike on closing price with the change 42.96%.
- On 19 December 2013 there is spike on closing price with the change 32.38%.
- On 20 July 2017 there is spike on closing price with the change 23.94%.

SQL Project Greencycles Online Movie Rental Shop Data Analysis Using PostgreSQL

Project Description

Problem :
Inventory Management
Ensuring the availability of popular movies and managing inventory effectively

Challenges :
Help the company operate big query and gain insight from data.

Project Result

[Click here to get full code]

Query Task

1. In the email system there was a problem with names where either the first name or the last name is more than 10 characters long. Find these customers and output the list of these first and last names in all lower case

first_names text	last_names text	email text
william	satterfield	william.satterfield@sakilacustomer.org
christopher	greco	christopher.greco@sakilacustomer.org
henry	billingsley	henry.billingsley@sakilacustomer.org
roger	quintanilla	roger.quintanilla@sakilacustomer.org
jonathan	scarborough	jonathan.scarborough@sakilacustomer.org

2. Extract the last 5 characters of the email address first. The email address always ends with '.org'. How can you extract just the dot '.' from the email address?

right text	left text
r.org	.
r.org	.
r.org	.
r.org	.
r.org	.

3. You need to create an anonymized version of the email addresses. It should be the first character followed by '***' and then the last part starting with '@'.

anonymized_email text
M***@sakilacustomer.org
P***@sakilacustomer.org
L***@sakilacustomer.org
B***@sakilacustomer.org
E***@sakilacustomer.org

4. In this challenge you have only the email address and the last name of the customers. You need to extract the first name from the email address and concatenate it with the last name. It should be in the form: "Last name, First name".

email text	position integer	left text	?column? text
MARY.SMITH@sakilacustomer.org	6	MARY	SMITH,MARY
PATRICIA.JOHNSON@sakilacustomer.org	10	PATRICIA	JOHNSON,PATRICIA

5. You need to create an anonymized form of the email addresses

email text	?column? text
MARY.SMITH@sakilacustomer.org	M***.S***@sakilacustomer.org
PATRICIA.JOHNSON@sakilacustomer.org	P***.J***@sakilacustomer.org
LINDA.WILLIAMS@sakilacustomer.org	L***.W***@sakilacustomer.org
BARBARA.JONES@sakilacustomer.org	B***.J***@sakilacustomer.org

6. What's the highest amount one customer has spent in a week?

month numeric	total_payment_amount numeric
4	27226.52
3	23886.56
2	9745.61
1	4710.70

7. You need to sum payments and group in the following formats

total_amount numeric	day text
39.91	Thu,03:45
16.96	Thu,10:05
13.96	Mon,05:30
10.98	Sat,11:20
2.99	Fri,07:01

8. You need to create a list for the suppcity team of all rental durations of customer with customer_id 35. Also you need to find out for the suppcity team which customer has the longest average rental duration?

customer_id smallint	avg interval
315	6 days 14:13:22.5
187	5 days 34:58:38.571428
321	5 days 32:56:32.727273
539	5 days 31:39:57.272727
436	5 days 31:09:46

9. Your manager is thinking about increasing the prices for films that are more expensive to replace. Create a list of the films including the relation of rental rate where the rental rate is less than 4% of the replacement cost.

film_id [PK] integer	percentage numeric
417	3.30
663	3.30
52	3.30
163	3.30
733	3.30

Big Data Analytics Project Salicyl Sales Dashboad on Kimia Farma

Project Description

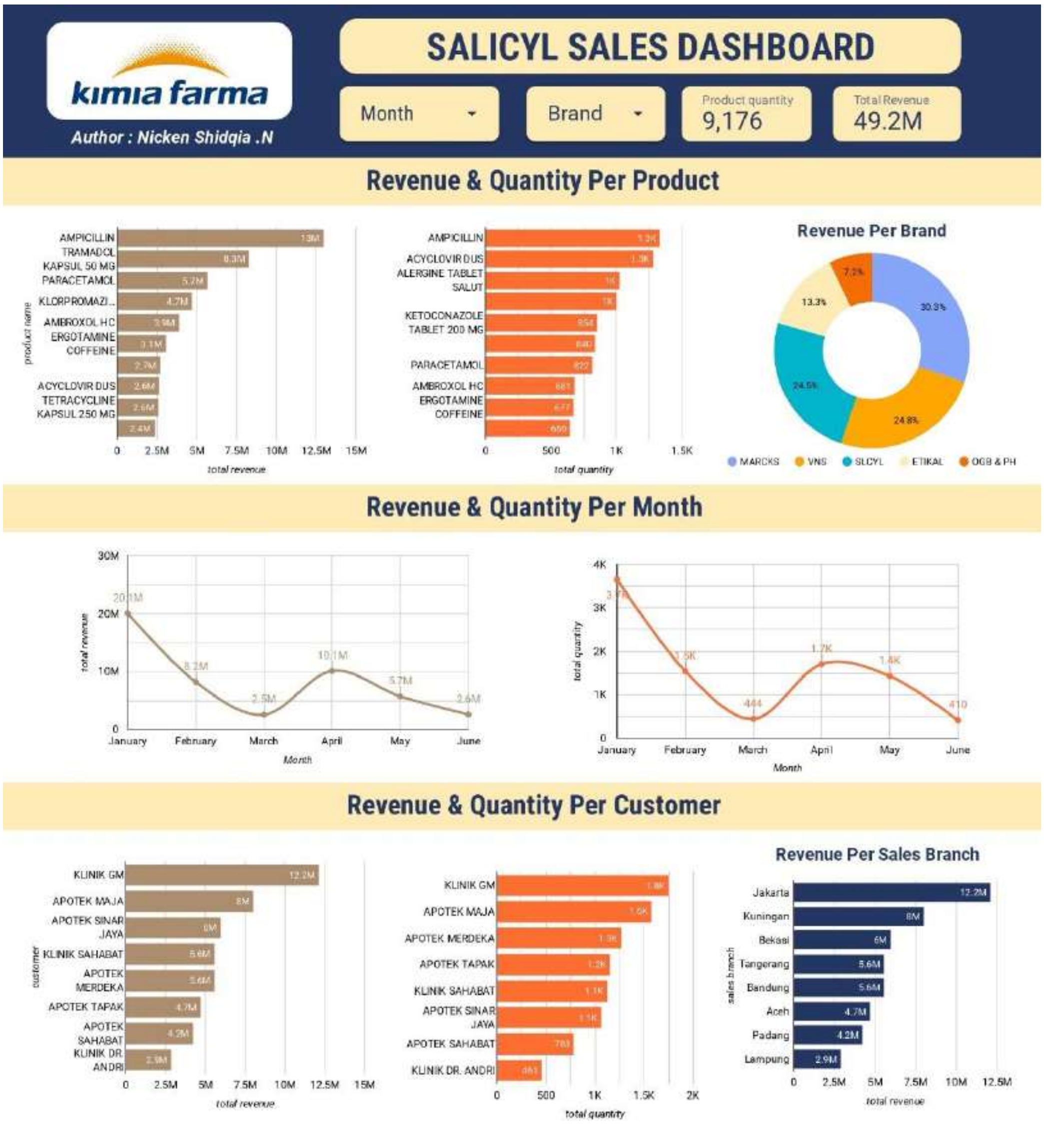
Challenges :

- Create a datamart design (Consisting of base tables and aggregate tables)
- Create a salicycl sales data visualization dashboard
- Create insights and provide additional complementary data

Project Result

[Click here to get dashboard link]

Dashboard Visualization



Insight

All Kimia Farma Brand Sales

- The highest revenue based on product category is Ampicillin with 13 M and total quantity 1.3K
- The highest revenue based on brand category is Marcks with 30.3%, followed by VNS 24.8%, and SLCYL 24.5%.
- The highest revenue based on sales branch category is Jakarta with 12.2 M.
- Sales of Kimia Farma are fluctuating with the highest revenue is happened on January 2022 with 20.1 M, while the lowest revenue is happened on March 2022 with 2.5 M
- The highest revenue based on customer category is Klinik GM 12.2 M with total quantity 1.8K.

Salicyl Brand Sales

- The highest revenue based on product category is Paracetamol with 5.7 M and total quantity 840.
- The highest revenue based on sales branch category is Jakarta with 5.6 M.
- Total revenue of sales Salicyl product is 12 M with total quantity 1,892.
- The highest revenue based on customer category is Klinik GM 5.6 M with total quantity 880.

Additional Complementary Data

Geographic Information:

- Latitude and longitude of each distributor's and brach location
- City, state, or region where distributors and brach are located.

Promotional Activities:

- Promotion Type, example discounts, bundle offers, seasonal promotions.
- Promotion Duration : Start and end dates for each promotional activity.
- Promotion Channels: Where the promotions are advertised or offered (in-store, online, specific platforms).

Competitor Data:

- Competitor Product Information
- Competitor Pricing
- Market Share
- Promotional Strategies
- Customer Reviews and Feedback

Startup Venture Funding Dashboard Data Analysis

Project Description

Overview :

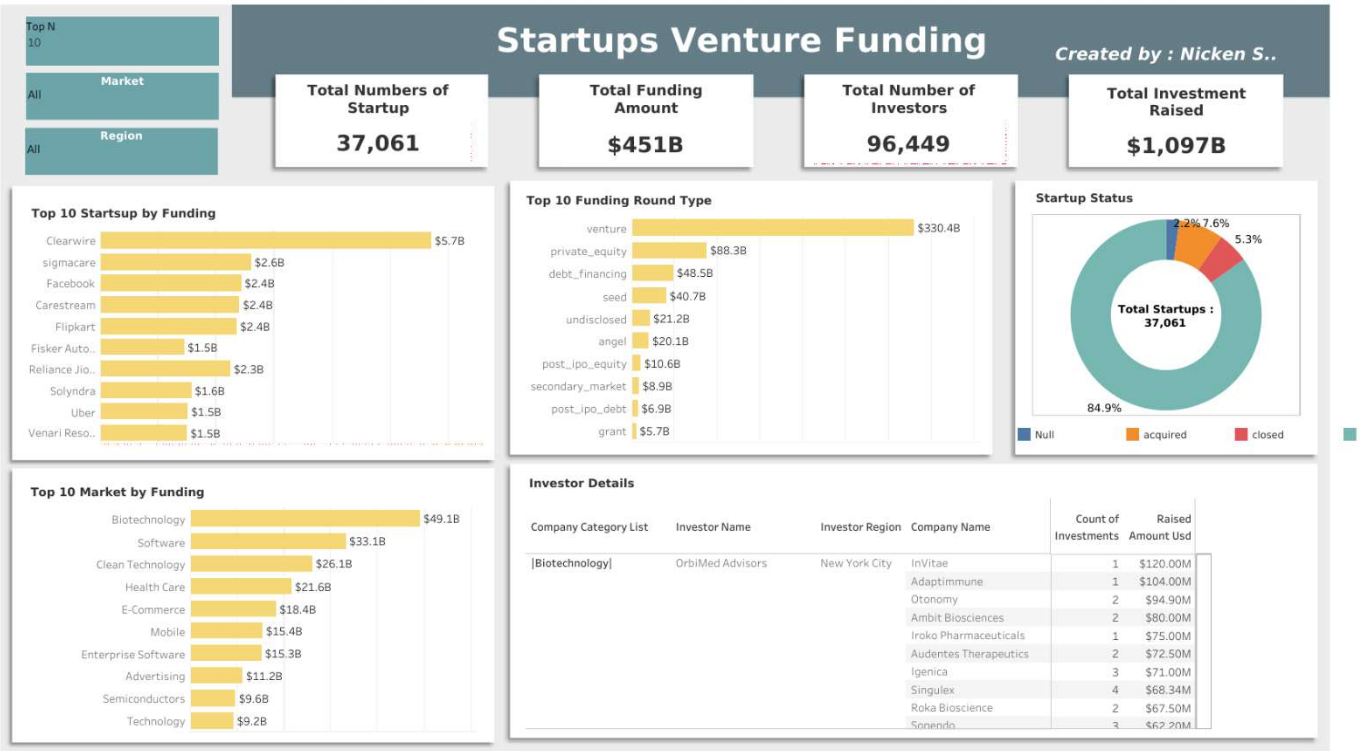
The Startup Venture Funding Dashboard is a comprehensive visual representation of the dynamic landscape of startup funding, providing valuable insights into the top startups, funding round types, markets, startup statuses, and investor details. The dashboard utilizes five key sheets to present a holistic view of the startup ecosystem.

Project Goal

The project aims to provide a centralized and visually intuitive platform that encapsulates key metrics and insights regarding startup funding.

Project Result

[Click here to get dashboard link]



Dashboard Insight:

1. Total Number of Startups: 37,061
2. Total Funding Amount: $451 billion
3. Total Number of Investors: 96,449

4. Total Investment Raised: $1,097 billion
5. Top 10 Startups by Funding:
    - Bar chart displaying the funding amounts of the top 10 startups.
    - Key Insight: Clearwire leads with $5.7 billion in funding.
6. Top 10 Funding Round Types:
    - Bar chart illustrating the distribution of funding across different round types.
    - Key Insight: Venture rounds dominate with a total funding of $330.4 billion.
7. Top 10 Markets by Funding:
    - Bar chart showcasing funding amounts in various markets.
    - Key Insight: Biotechnology emerges as the leading market with $49.1 billion in funding.
8. Startup Status:
    - Donut pie chart representing the distribution of startups based on their status.
    - Key Insight: 84.9% of startups are currently operating.
9. Investor Details:
    - Sheet providing detailed information on investors participating in startup funding.

Recommendation

- Given that venture rounds dominate with $330.4 billion, investors may consider maintaining a focus on venture funding as it represents a significant portion of the overall funding landscape.
- Biotechnology emerges as the leading market with $49.1 billion in funding. Entrepreneurs and investors may explore opportunities within the biotechnology sector, considering the demonstrated investor interest and potential for growth.
- As 84.9% of startups are currently operating, there is a strong emphasis on sustaining and growing existing ventures. Investors may focus on startups with a proven track record of operation for potential long-term returns.
- Startups with high funding amounts, such as Clearwire with $5.7 billion, may be attractive for potential strategic partnerships. Investors and corporations could explore collaboration opportunities with these high-performing startups.