CASE 2

BAN 620 - DATA MINING

Group 6

Sarath kumar Vatyam - SS2405

Abhi Ram Sai Lakkavarapu - BD2500

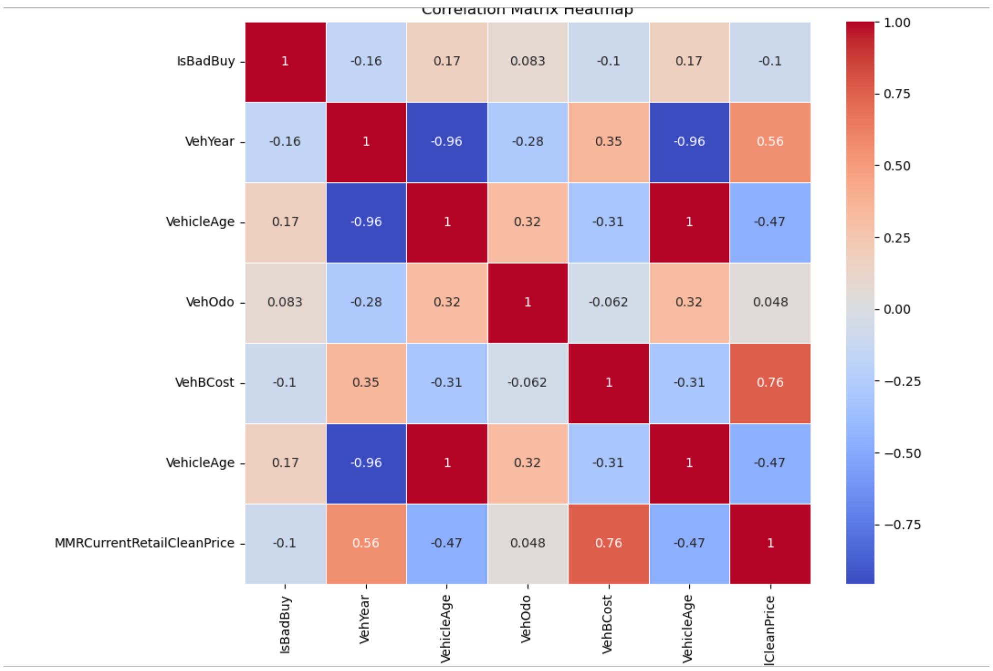
Neeharika Vijjada – AS1195

This case study examines Carvana, a retailer specializing in used cars that aimed to revolutionize the car buying process. Carvana was established in 2012. Launched its services in 2013 allowing customers to research, finance and purchase owned vehicles entirely through their website. The company stood out by introducing features such, as car vending machines and a 7-day money back guarantee.

Before its launch in 2011 Carvana organized a Kaggle competition where data scientists were invited to develop an algorithm of predicting whether cars purchased at auctions would have undisclosed issues. A total of 570 teams. Carvana rewarded the winning team with $10,000.

Although Carvana went public in 2017 it experienced a drop in share value on its day of trading. Some raised questions about whether the company's data science and algorithms provided an advantage for its growth or if it was more of a marketing strategy. Additionally, this case study explores the connections between Carvana's founders, its parent company Drive Time and how these relationships impacted the business.

Overall, this case study offers insights, into Carvana's business model and its efforts to leverage data science as a means of disrupting an industry that's both vast and fragmented.



From the Carvana dataset, we've tried to find out the correlation between different numerical factors and drew conclusions whether it is a bad buy or not. In the dataset "isBadBuy" column determines whether it's a bad buy or good buy when it comes to a car purchase. From the heatmap we can see that the "isBadBuy" is effected by all the other elements in different ratios.

The decision to buy the vehicle is highly effected by Vehicle age. Logically the older the vehicle the lesser the efficiency and the life span.

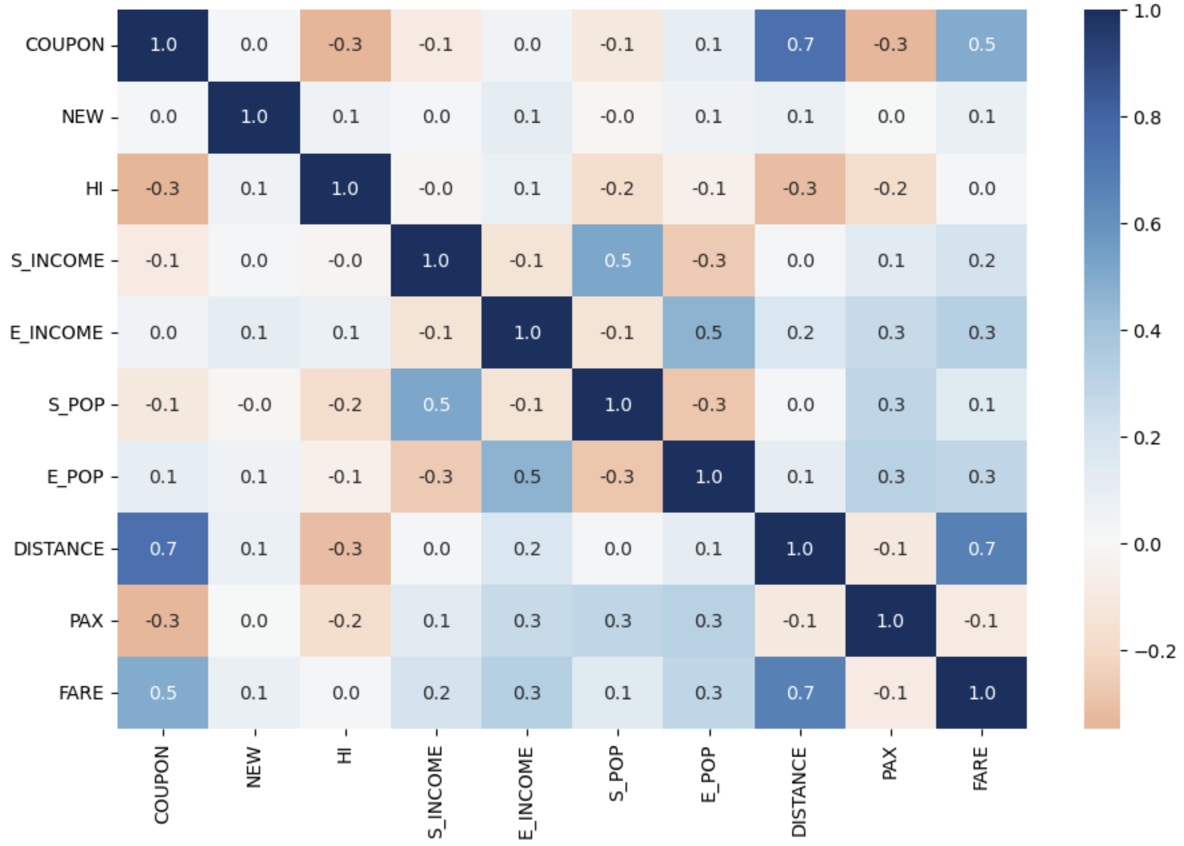
AIRFARES CASE

Given the dataset named “Airfares”, contains multiple variables, and has the data of fares between various cities and different airlines. It also includes the number of passengers travel between the given cities and number of layovers they take and whether they travel in vacation season or not.

Based on this data, we can build an appropriate model that can predict the fares between given routes.

As per the data, the airlines which are starting from Minneapolis have the highest average fare and the fare for Portland, Oregon has lowest average fare.

As Fare is the key business variable , we can check the correlation between Fare and other variables to see how they are correlated each other as per below Heat Map



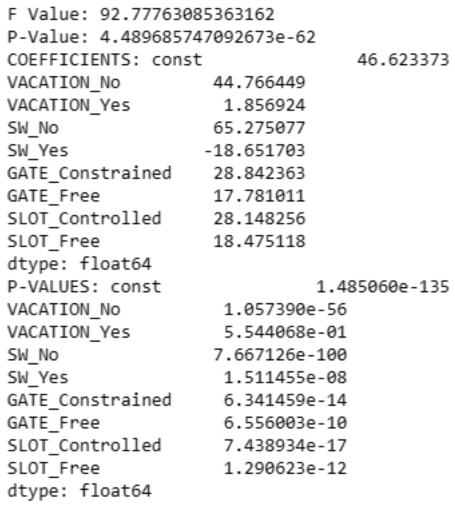
	COUPON	NEW	HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX	FARE
COUPON	1.0	0.0	-0.3	-0.1	0.0	-0.1	0.1	0.7	-0.3	0.5
NEW	0.0	1.0	0.1	0.0	0.1	-0.0	0.1	0.1	0.0	0.1
HI	-0.3	0.1	1.0	-0.0	0.1	-0.2	-0.1	-0.3	-0.2	0.0
S_INCOME	-0.1	0.0	-0.0	1.0	-0.1	0.5	-0.3	0.0	0.1	0.2
E_INCOME	0.0	0.1	0.1	-0.1	1.0	-0.1	0.5	0.2	0.3	0.3
S_POP	-0.1	-0.0	-0.2	0.5	-0.1	1.0	-0.3	0.0	0.3	0.1
E_POP	0.1	0.1	-0.1	-0.3	0.5	-0.3	1.0	0.1	0.3	0.3
DISTANCE	0.7	0.1	-0.3	0.0	0.2	0.0	0.1	1.0	-0.1	0.7
PAX	-0.3	0.0	-0.2	0.1	0.3	0.3	0.3	-0.1	1.0	-0.1
FARE	0.5	0.1	0.0	0.2	0.3	0.1	0.3	0.7	-0.1	1.0

We see that FARE is highly correlated with the variables Coupon(0.5) and Distance(0.7) , thus these can be key variables for the change in FARE.

We observe that HI is not correlated with FARE , we can exclude the variable for the relative analysis.

We can conclude that FARE and distance has directly proportionate with the FARE , as distance increases , FARE also increases.

We looked at several factors like "Holiday," "South West (SW) Airlines," "Gate," and "Slot" from our dataset. We wanted to see if these factors have any influence on airfare ("FARE"). To do this, we transformed these factors into a format that our analysis could understand and built a model to see if they matter.



```
F Value: 92.77763085363162
P-Value: 4.489685747092673e-62
COEFFICIENTS: const               46.623373
VACATION_No         44.766449
VACATION_Yes         1.856924
SW_No               65.275077
SW_Yes             -18.651703
GATE_Constrained    28.842363
GATE_Free           17.781011
SLOT_Controlled     28.148256
SLOT_Free           18.475118
dtype: float64
P-VALUES: const               1.485060e-135
VACATION_No          1.057390e-56
VACATION_Yes         5.544068e-01
SW_No               7.667126e-100
SW_Yes               1.511455e-08
GATE_Constrained     6.341459e-14
GATE_Free            6.556003e-10
SLOT_Controlled      7.438934e-17
SLOT_Free            1.290623e-12
dtype: float64
```

The Impact of South West (SW) Airlines: Among these factors, we discovered that when a route is served by South West (SW) Airlines, it tends to affect the airfare. Specifically, if South West is operating on a route, there's a 43% chance that the airfare will be different compared to routes where South West isn't operating. This is shown by the 43% coefficient.

Gate and Slot Influence: On the other hand, the numbers for "Gate" (3.2) and "Slot" (3.86) suggest that airfare isn't significantly influenced by how crowded or controlled the destination airport is. In simpler terms, whether the airport is super busy or not doesn't seem to affect airfare much.

Statistical Significance: Some factors, like "No Holiday," "No South West (SW)," "Constrained Gate," and "Controlled Slot," have very low p-values. This means they are statistically important, and they do have an impact on airfare.

In a nutshell, our analysis showed that having South West Airlines on a route makes a real difference in airfare, while how busy an airport is (Gate and Slot) doesn't seem to matter much.

A model has been developed to predict the fares.

This model includes all the variables from the given dataset and the categorical variables have been made into numerical by using encoding techniques(creating dummies using pandas) and these dummy variables has been copied into source data, removing the original categorical objective data.

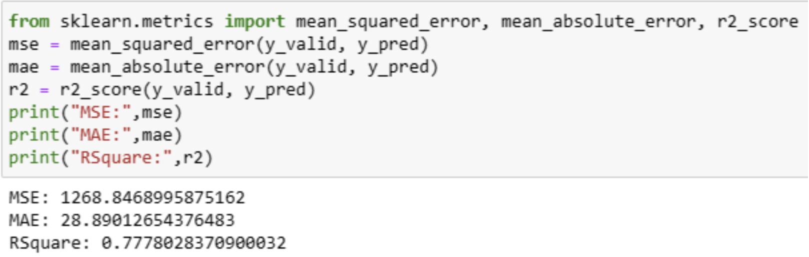
As the data was suitable for model, we have split the data into two

1. Feature variables
2. Outcome Variable (FARE)

The model has been trained, validated and tested using the python library scikit learn and we have considered the linear regression model for this prediction as the outcome variable is continuous.

The trained model has been tested for accuracy using regression statistics i.e, R square, Mean square error and Mean Absolute error.

As per the R-square score of ~77% , the model is moderately accurate in predicting the fare in respect to the selected feature variables.



```
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
mse = mean_squared_error(y_valid, y_pred)
mae = mean_absolute_error(y_valid, y_pred)
r2 = r2_score(y_valid, y_pred)
print("MSE:",mse)
print("MAE:",mae)
print("RSquare:",r2)
```

```
MSE: 1268.8468995875162
MAE: 28.89012654376483
RSquare: 0.7778028370900032
```

For predicting more accurate FARES, some feature variables which might not relate to the target variable i.e., FARE can be removed from data to fine tune the model to get accurate fares and this regression model can be fed with more data unlike the current data of only 628 fields. With more training data, models can be better trained and validated to get accurate results.

Also, scaling can be performed to feed the model with less errors and better performance.

The model that we have built, suits if new Airport has brought into service ,

- Model has built using generalized variables taking business into account
- As we see there are not many outliers in the variables , we can be firm that if the new airport data follow general trends , the current model would be helpful for FARE predictions.

The model may not work efficiently because of following reasons –

- Lack of variables data for the new airport
- Inappropriate market conditions (Seasonal , Geo- political and Economic factors)
- FARE might change due to Competitive reactions due to increase of other airlines operating in the same region.