</talentlabs>

Assignment 9 - Multivariate Analysis (29 points)

Instructions

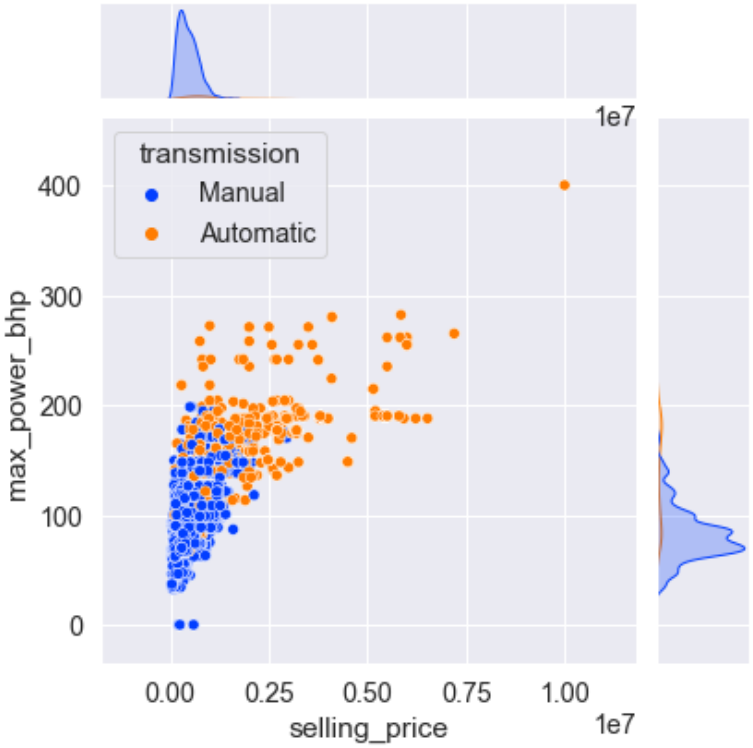
1. Answer the below question in the boxes if needed.
2. For coding exercises, code in a single google colab notebook and zip all your code before submission.
3. Please submit the assignment through TalentLabs Learning System

</talentlabs>

Question 1 (5 points)

Questions are based on automobile characteristics data. (dataset not required for these questions)

(*Note: max power is measured in horse power, Brake horsepower or bhp refers to the horsepower of the car after taking into consideration friction between a car's tyres and the road, selling price is measured in dollars, 1e7 is 10,000,000, so a 0.75 means 0.75 x 10,000,000 = 7 million and 500k dollars*)



Based on the plot above, answer the following questions:

1. What kind of a plot is this? What kind of variables are plotted here, name them and their types ? (3 points)

Scatterplot 1. Max Power (Numerical Variable - Continuous) 2. Selling Price (Numerical Variable – Continuous)

2. Do you find any findings in the chart? Give 1 insights based on the chart (1 point)

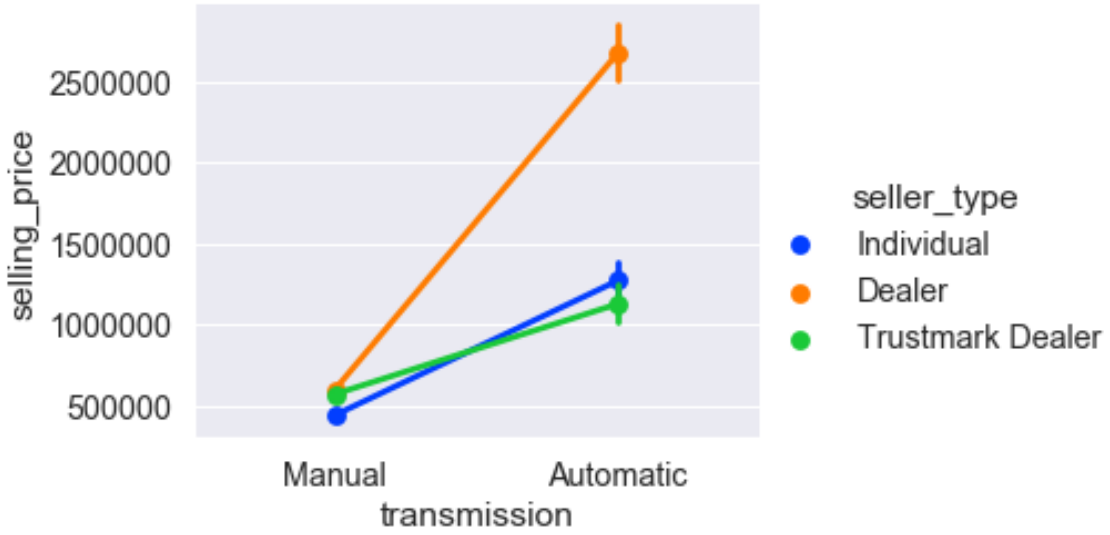
Higher max power might correspond to higher selling prices, indicating a positive correlation.

3. Do you see any outliers in the chart? (1 point)

No, there is one that is far away, but it is relatable as higher bhp has higher selling price.

</talentlabs>

Question 2 (4 points)



Given plot 2 answer the following (assume the points show an average):

1. What kind of a plot is this? What kind of variables are plotted here, name them and their types ? (2 points)

Line Plot

- Transmission Type (Categorical Variable)
- Selling Price (Numerical Variable - Continuous)
- Seller Type (Categorical Variable)

2. Do you see any difference based on seller types? If yes, what do you see here? (2 points)

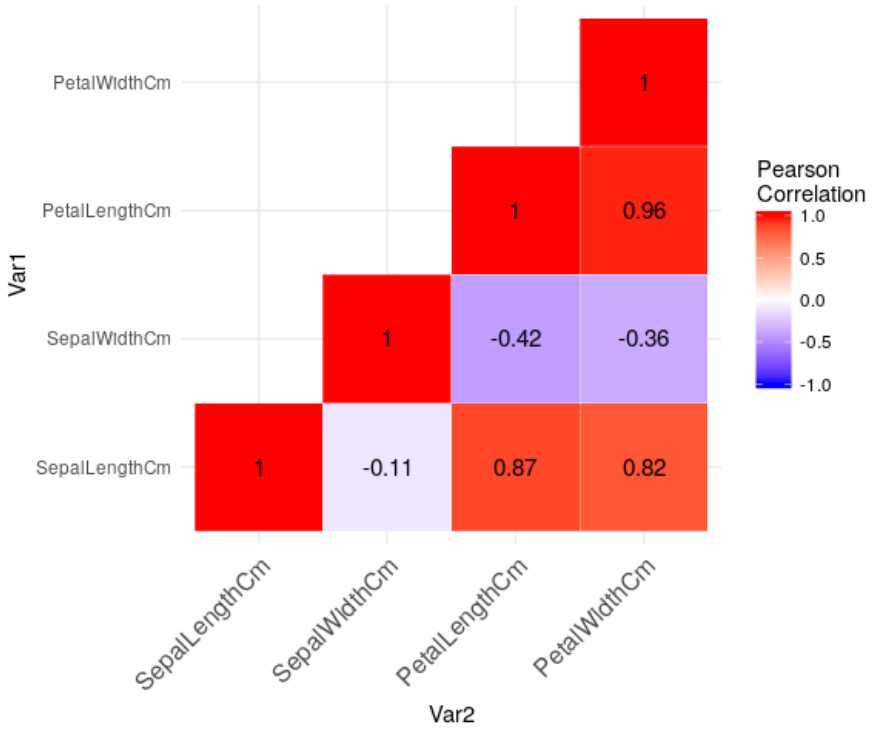
Seller Type Influence on Price:

- Dealers generally sell vehicles at a much higher price than both Individual sellers and Trustmark Dealers, especially for automatic transmission cars.
- The price difference for Individual sellers and Trustmark Dealers is minimal, indicating that these seller types may have similar pricing strategies or market positioning.

</talentlabs>

Question 3 (4 points)

Questions are based on the Iris species data (https://www.kaggle.com/datasets/uciml/iris)



Given the plot above, answer the following:

1. What kind of a plot is this? What kind of variables are plotted here, name them and their types ? (2 points)

Heatmap

- Sepal Length (Numerical Variable - Continuous)
- Sepal Width (Numerical Variable - Continuous)
- Petal Length (Numerical Variable - Continuous)
- Petal Width (Numerical Variable - Continuous)

2 .What insights can you draw from here regarding the relationships between the variables? Give 2 insights here. (2 points)

1. There is a very strong positive correlation between Petal Length and Petal Width (correlation coefficient = 0.96). This suggests that as the petal length increases, the petal width also tends to increase proportionally.

2. There are also strong positive correlations between Sepal Length and Petal Length (0.87) and between Sepal Length and Petal Width (0.82). This indicates that flowers with longer sepals generally also have longer and wider petals.

</talentlabs>

Question 4 (16 points)
Note: Please submit the Google Colab or Jupyter Notebook for this question.

Load the titanic dataset using seaborn given the code below and answer the questions below:

```
import seaborn as sns
df = sns.load_dataset('titanic');
```

Study the dataset and the goal here: https://www.kaggle.com/competitions/titanic.
You can use seaborn or matplotlib or both.

Data Dictionary

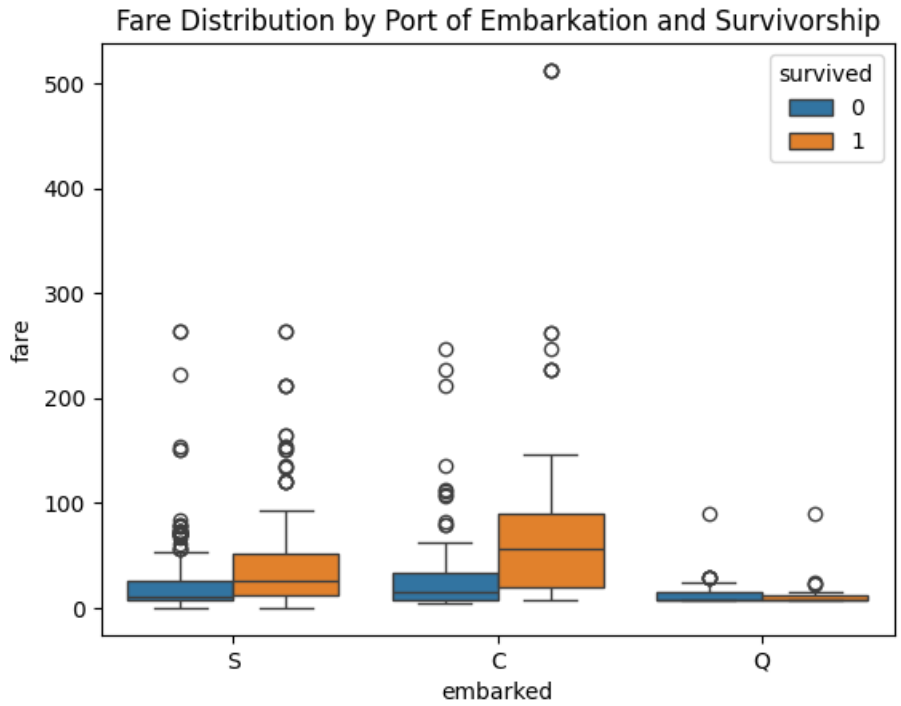
Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

1. Using a charting tool of your choice(bar or box or factor plots), show how port of embarkation and survivorship relate to fare in one plot! (use survived as color/hue)

   Write about queenstown and cherbourg fare rates, do you see any difference on an average for those who survived/not survived? (4 points)

Screenshot of the chart:

</talentlabs>

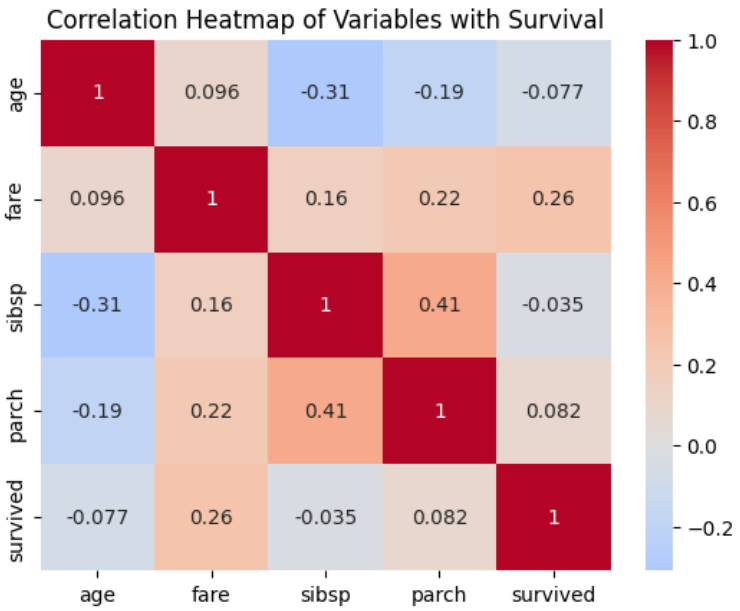


Insights:

- There is a significant difference in fare rates for those who survived versus those who did not, especially for passengers from Cherbourg, where higher fares are associated with a higher likelihood of survival. However, for passengers from Queenstown, the fare difference between survivors and non-survivors is minimal, indicating that other factors may have influenced survival rates for these passengers.

2. Correlate numerical variables(Age, Fare) and Discrete Variables (treat sibsp and parch as discrete variables) with survival (create variables survived and not survived) and show via a heatmap. Which two variables have the strongest relationship? Which variable has the strongest variable with those who survived? (4 points)

Screenshot of the chart:



Two variables with strongest relationship: parch & sibsp

Variable with strongest relationship with survival: fare

3. Create a pivot table using Survival and Sex on the index, port of embarkation on the columns and Average Fare and Counts as the metric/aggregation function, fill any missing values with 0's.

   What is the highest and lowest average fare in the table for those who survived and for those who didn't survive? Jot down if that person was a male or female and which port that person embarked from for each.

   (8 points)

Screenshot of the table:

		mean			count		
embarked		C	Q	S	C	Q	S
survived	sex						
0	female	16.215278	10.904633	25.728508	9	9	63
	male	38.065342	13.911732	19.881281	66	38	364
1	female	83.460286	13.211733	44.596518	64	27	140
	male	71.468545	12.916667	30.366286	29	3	77

</talentlabs>

Highest average fare for the ones who survived: 83.460286
Male or Female: Female
Port of Embarkation: Cherbourg

Highest average fare for the ones who did not survived: 38.065342
Male or Female: Male
Port of Embarkation: Cherbourg

Lowest average fare for the ones who survived: 12.916667
Male or Female: Male
Port of Embarkation: Queenstown

Lowest average fare for the ones who did not survived: 10.904633
Male or Female: Female
Port of Embarkation: Queenstown