# Assignment 9 - Multivariate Analysis (29 points)

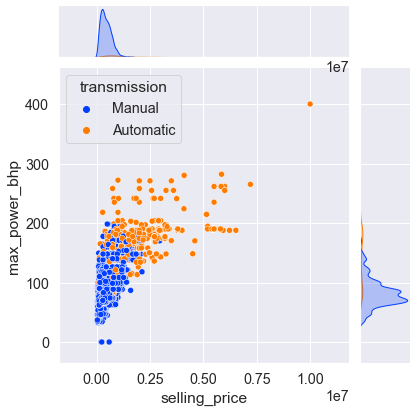
## Instructions

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

**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)

|  |
| --- |
| * Scatter plot. * They are both numeric variables, where the data type as below: * selling\_price: integer * max\_power\_bhp: integer |

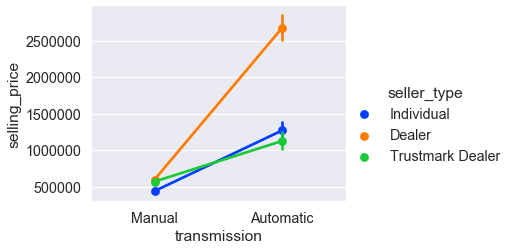
1. Do you find any findings in the chart? Give 1 insights based on the chart (1 point)

|  |
| --- |
| Automatic transmission cars have higher brake horsepower compared to manual transmission cars in general. |

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

|  |
| --- |
| Yes. There is an outlier where max\_power\_bhp = 400 and selling\_price = 1.00. |

**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)

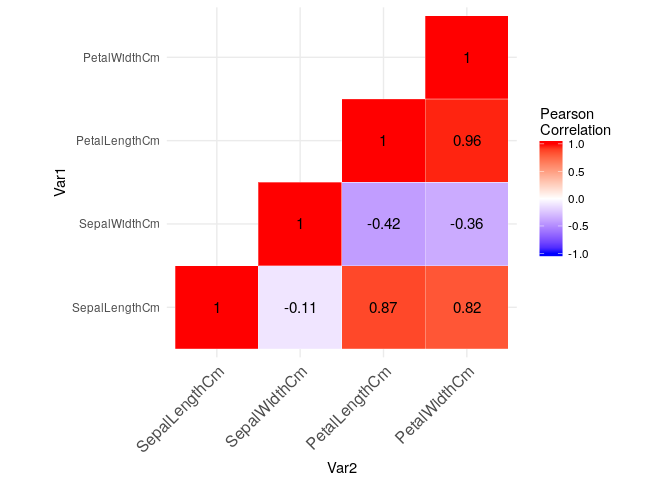
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| * Multivariate line chart  |  |  |  | | --- | --- | --- | | **Column** | **Kind of variables** | **Data type** | | selling\_price | Numerical | Integer | | transmission | Nominal | String - categorical | |

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

|  |
| --- |
| Yes.  The info we get:   * Dealer sold automatic transmission cars at the price that was at least twice as high as those offered by individual and Trustmark dealer. * Dealer sold automatic transmission cars at the price that was at least 5 times greater than the manual transmission cars. |

**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.   |  |  |  | | --- | --- | --- | | **Column** | **Kind of variables** | **Data types** | | SepalLengthCm | Numerical | float | | SepalWidthCm | Numerical | float | | PetalLengthCm | Numerical | float | | PetalWidthCm | Numerical | float | |

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

|  |
| --- |
| * Sepal Length is strongly correlated with Petal Width. * Sepal Length is strongly correlated with Petal Length. |

**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

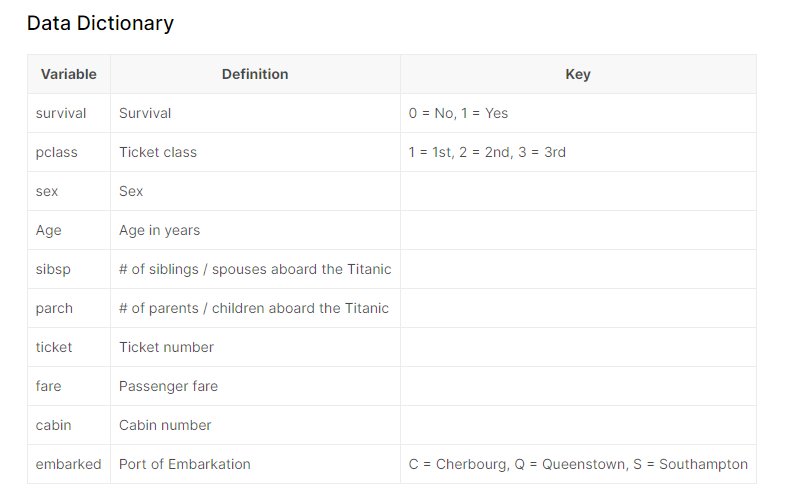
below:

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.



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:    Insights:   * **Queenstown’s fare rate**: No significant difference for both survivors and non-survivors. * **Cherbourg’s fare rate**: The survivors have a much higher fare rate, which are at least twice as high as non-survivors. |

1. 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: survival\_0, survival\_1  Variable with strongest relationship with survival: age |

1. 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:    Highest average fare for the ones who survived:  Male or Female: female  Port of Embarkation: Cherbourg  Highest average fare for the ones who did not survived:  Male or Female: male  Port of Embarkation: Cherbourg  Lowest average fare for the ones who survived:  Male or Female: male  Port of Embarkation: Queenstown  Lowest average fare for the ones who did not survived:  Male or Female: female  Port of Embarkation: Queenstown |