Mod 3. Exploratory Data Analysis

Preliminary step in data analysis to:

- ✓ Summarize main characteristics of the data
- ✓ Get better understanding of the dataset
- ✓ Uncover relationships between variables
- ✓ Extract important variables

- Descriptive Statistics: Describe basic features of a dataset and obtains a short summary about the sample
- Group By
- ANOVA: Analysis of variance. The variation in a set of observations is divided into distinct components
- Correlation
- Correlation - Statistics: Pearson correlation and Correlation Heatmaps

Descriptive Statistics

- Describe basic features of data
- Giving short summaries about the sample and measures of the data

*Summarize statistics

`df.describe()` ← Statistics for all numerical values

Any NaN values are automatically stripped in these statistics

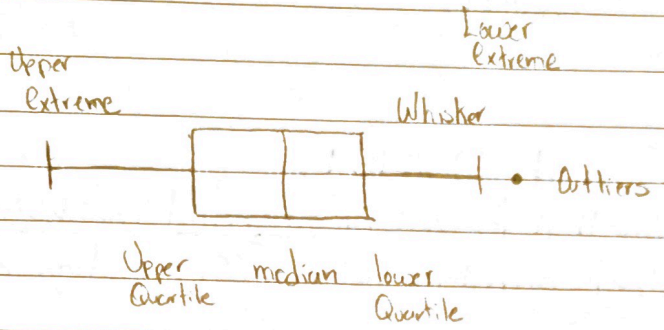
*Summarize the categorical data. Distribution of your different variables

`value_counts()` Variables that can be divided up into different categories or groups and have discrete values (# cars = int = engine)

```
drive_wheels_counts = df["drive-wheels"].value_counts()
drive_wheels_counts.rename(columns = {'drive-wheels':'value_counts'}, inplace = True)
drive_wheels_counts.index.name = 'drive-wheels'
```

drive-wheels	value_counts
fwd	118
rwd	75
4wd	8

Box Plots



Easy to compare between groups

```
sns.boxplot (x = "drive-wheels', y = 'price', data = df)
```

Scatter Plot

- Each observation represented as a point
- Scatter plot show the relationship between two variables
  1. Predictor / independent variables on x-axis
  2. Target / dependent variables on y-axis

```
y = df ["engine-size"]
x = df ["price"]
plt. scatter (x,y)

plt. title ("Scatterplot of Engine Size vs Price")
plt. xlabel ("Engine Size")
plt. ylabel ("Price")
```

Positive linear relationship between these two variables

Group by () - Example

Grouping Data

- Use Panda df. groupby () method:
  The group by method is used on categorical variables, groups the data into subsets according to the different categories of that variable.
- You can group by a single variable or you can group by multiple variables by passing in multiple variable names.

```
df_test = df [['drive-wheels', 'body-style', 'price']]
df_grp = df_test.groupby (['drive-wheels', 'body-style'], as_index = False).mean()
df_grp
```

Agrega la 3ra columna de precio

	drive-wheels	body-style	price
0	4wd	hatchback	7603.0
1	4wd	Sedan	12647.00

grouped int subcategories

Transform it to a pivot table to make it easy to read

Pandas method - Pivot()

One variable displayed along the columns and the other variable displayed along the rows → Similar to excel

fila (row)

```
df_pivot = df_grp.pivot (index = 'drive_wheels', columns = 'body_style')
```

	Price		
Body-style	Convertible	hardtop	hatchback
drive-wheels			
4wd	2039.-	20239.-	7603.-

Norma

Heatmap

Pivot tables can be converted to heatmaps

- Plot target variable over multiple variables

```
plt.pcolor(df_pivot, cmap='RdBu')
plt.colorbar()
plt.show()
```

Analysis of variance ANOVA

- Statistical comparison of groups

Example: average price of different vehicle makes.

ANalysis Of Variance (ANOVA)

* Why do we perform ANOVA?

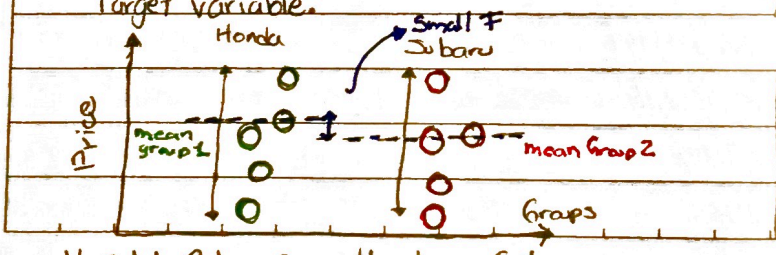
Finding correlation between different groups of a categorical variable.

* What we obtain from ANOVA?

- F-test score: variation between sample group means divided by variation within sample group
- p-value: confidence degree

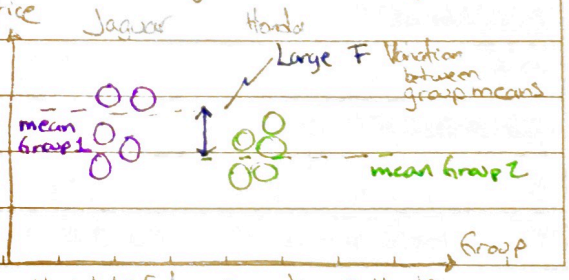
F-test - calculates the ratio of variation between group means over the variation within each of the sample group means

- Small F imply poor correlation between variable categories and target variable.



Variable Categories: Honda y Subaru
Target Variable: Price

- Large F imply strong correlation between variable categories and target variable



Variable Categories: Jaguar, Honda
Target Variable: Price

Norma

Small F score between Hondas and Subarus because there's a small difference between the average prices → poor correlation

Large F value between Hondas and Jaguars because the difference between prices is very significant → strong correlation

- ANOVA between "Honda" and "Subaru"

```
df_anova = df[["make", "price"]]
grouped_anova = df_anova.groupby(["make"])
anova_results = stats.f_oneway(grouped_anova.get_group("honda")["price"], grouped_anova.get_group("subaru")["price"])
```

Anova results: F=0.1974403, p=F_oneway result(statistics=0.1974403, pvalue=0.660)

- ANOVA between "Honda" and "Jaguar"

Anova results: F=400.9258, p=F_oneway result(statistics=400.925, p value=1.055)

p-value is larger than 0.05

Strong correlation between categorical variable and other variables if ANOVA test gives a large F test and a small ?? p-value * check

Correlation

What is correlation?

Measures to what extent different variables are interdependent

- Examples:
  - Lung cancer -> Smoking
  - Rain -> Umbrella

Correlation doesn't imply causation : Umbrella and rain are correlated, but we would not have enough information to say whether the umbrella caused the rain or the rain caused the umbrella.

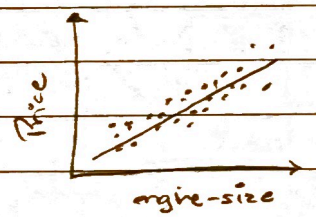
Correlation - Positive Linear Relationship

- Correlation between two features (engine-size and price).

```
sns.regplot(x="engine-size", y="prices", data=df)
plt.ylim(0,)
```

The main goal of this plot is to see whether the "engine-size" has any impact on the price

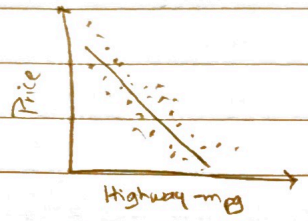
Positive → linear relationship between two variables ∴ Positive correlation between engine size & price



Correlation - Negative Linear Regression

- Correlation between two features (highway-mpg and price).

```
sns.regplot(x="highway-mpg", y="price", data=df)
plt.ylim(0,)
```

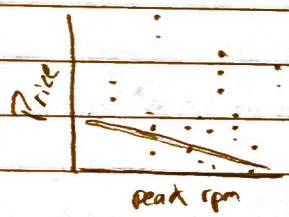


Negative linear relationship between highway-mpg and price

Correlation - Weak correlation

- Weak correlation between two features (peak-rpm and price)

```
sns.regplot(x="peak-rpm", y="prices", data=df)
plt.ylim(0,)
```



We can't use peak rpm to predict price

CORRELATION STATISTICS

* Pearson Correlation

- Measure the strength of the correlation between two features
  - Correlation Coefficient
  - P-value

- Correlation Coefficient
  - Close to +1 : Large Positive relationship
  - " -1 : Large Negative relationship
  - " 0 : No relationship

- P-Value
  - P-value < 0.001 Strong Certainty in the result
  - " < 0.05 Moderate "
  - " < 0.1 Weak "
  - " > 0.1 No "

- Strong Correlation:

Correlation coefficient close to 1 or -1

P-value less than 0.001

```
pearson_coef, p_value = stats.pearsonr(df['horsepower'], df['price'])
```

- Pearson Correlation: 0.81
- P-value: 9.35e-48

Correlation coefficient is close to 1

P value is much smaller than 0.001 so we can conclude there's a strong correlation