1. Assignment-based subjective questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer:

- Rental bikes are taken more on **Falls** compare to other seasons and low on **Spring** season
- For both working day and non-working day
- In mid-year [between May to October] we have high rental bikes
- There are more rental bikes on 2019 compare to 2018
- Holidays has fewer rental bikes compare to non-holiday
- For all weekdays, rental bikes have same median value
- During clear day we can see more rental bikes and there are no rental bikes taken during Heavy rain

2. **Why is it important to use drop_first=True during dummy variable creation?**

Answer:

If we have **N** number of dummy values, we can explain these **N** number of dummy value using **N-1** variables.

Drop_first = True is important as this reduces number of features in the dataset as well as correlation between the features.

E.g., Consider we have seasons has 4 values [Spring, Falls, Summer, Winter], so we have 4 dummy variables and these 4 seasons can be expressed using only 3 dummy variables [N=4, N-1=3].

Let's drop Spring here and we have Falls, Summer & Winter in sequence and below is how it expressed.

000 -> Spring

100 -> Falls

010 -> Summer

001 -> Winter

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Answer:

'temp' variable has the highest correlation with target variable among the numerical variables.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:

There are 4 assumptions in Linear Regression

i. **Linear relation between dependent and independent variables**: This is being tested using scatter plot in the initial to check whether there is any linear relation between the dependent and independent variables
ii. **Error term should be normally distributed with mean as 0:** Post completion of the model, I have plot the Distplot to see the residual distribution
iii. **Low or No Multicollinearity**: I have checked the VIF values of the features post model building and having the threshold value of 5 for the model, dropping the features one-by-on if a feature exceeds 5.
iv. **Homoscedasticity**: By checking if there are any pattern available in the scatter plot between y_train_pred and residual.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Answer:

Below are the top 3 features of the model,

- Light Snow [-0.3]
- Spring [-0.26]
- Year [0.25]

General Subjective Questions

1. **Explain the linear regression algorithm in detail**.

Answer:

Linear Regression is supervised learning method of machine learning.

Below are the steps in linear regression,

- Reading the data
  - This is where we read the data from source and store in a data-frame
- Data preparation
  - This is where we do data cleaning as below
    - Dropping the unnecessary columns
    - Changing the datatype
- Visualising the data
  - Plotting a Pairplot for all numerical variables to check if there is any linear relation between the dependent & independent
  - Plotting the boxplot between categorical variable vs dependent variable
- Create dummy variable
  - For categorical variables, creating dummy variables. For N variables, creating N-1 dummy variables

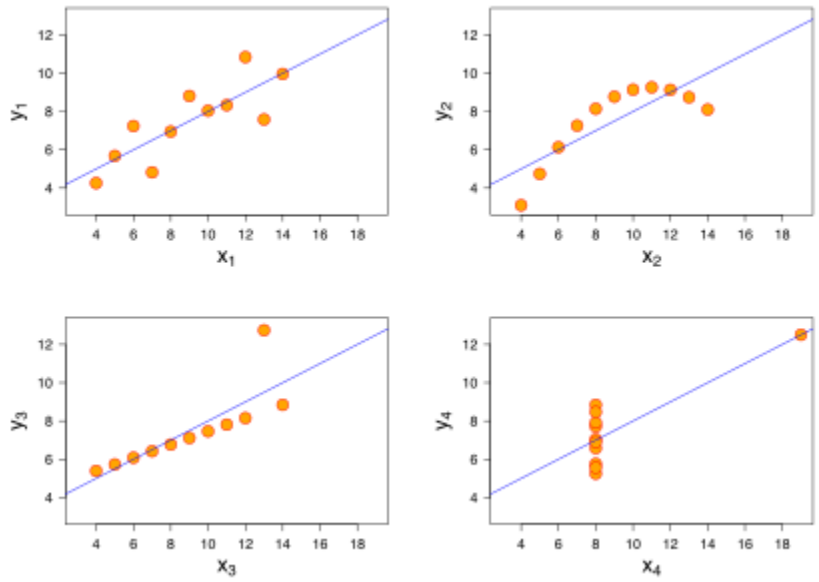
- Splitting data in training and testing data set
  - Splitting the data set either 70-30 or 80-20 for training and testing
  - Scaling the features for better interpretation. We have two types of scaling,
    - Min-Max Scaling [Normalization]
    - Standardisation
- Dividing the training dataset to X_train & y_train
- Building the Linear model
  - Identify the features to build the model based on correlation between independent and dependent
  - We can use RFE to identify the features in an automated way
  - Building the model with the features
    - We can add features step-by-step
    - We can add all features and drop one-by-one
  - Check the stats of the model
    - Check P-values of the features
    - R-square of the model
    - VIF of the features
  - If we add all the features, then drop the features with high P-value [more than 0.05]
    - Then build the model again with remaining features
    - Check the P-values and VIF for the features
    - If feature have high VIF [more than 5]
    - Drop the feature and build the model with remaining features
    - Repeat the same till the model is stable
  - If we add feature one-by-one,
    - Build a model with a dependent and 1 independent variable [identify using heatmap of correlation between features]
    - Check the model R-square value and P-values
    - Add another features to the model and check the R-square & P-value
    - If there is no significant increase in the R-square, drop the feature else proceed with add another feature
    - Repeat the process until the model is stable and not required to add new feature
- Residual analysis of the train set
  - Now get the y_train predict value
  - Plot error term to check if error term is normally distributed or not
  - Check if model is Homoscedasticity or not by plotting a scatter plot between y_train_pred vs residuals
- Making predictions to the final model
  - Scaling the test set [only transforming the data]
  - Dividing the X_test and y_test
  - Building the model with the features as the final model
- Model evaluation
  - Plotting scatter plot between y_test & y_pred
  - Checking the R-square value y_test and y_pred

- R-square for y_test & y_pred and model R-square value should be approximately equal with variance of 5

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted the scatter plot as mentioned below.



Each dataset contains 11 data points of (x, y) as below.

Anscombe's quartet dataset

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84

11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

For the above data set we can observe that all the four data set has the same mathematical stats. i.e.,

mean of X = 9

standard deviation of X=3.32

mean of Y = 7.5

standard deviation of Y=3.32

correlation between X & Y = 0.816

Observation:

i. Top-Left Scatter plot: It shows the there is a linear relation between X & Y
ii. Top-Right Scatter plot: values are normally distributed and not a linear relation
iii. Bottom-Left Scatter plot: Except for last point, remaining points has perfect linear relation
iv. Bottom-Right Scatter plot: There is no linear relation between X & Y, but has a high correlation at X=8

3. What is Pearson's R?

Correlation is a statistic that measures the relationship between two. It shows the strength of the relationship between the two variables as well as the direction and is represented numerically by the correlation coefficient. The numerical values of the correlation coefficient lies between -1.0 and +1.0.

- Pearson's Correlation Coefficient is also referred to as Pearson's R, the **Pearson product-moment correlation coefficient (PPMCC),** or bivariate correlation.
- Pearson's R is the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- Pearson's R cannot show the non-linear relationship between variables, and cannot differentiate between independent and dependent variables.
- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations
- There are certain requirements for Pearson's Correlation Coefficient.
    - Scale of measurement should be interval or ratio
    - Variables should be approximately normally distributed
    - Variables should have linear relation
    - There shouldn't be any outliers in data

Formula for Pearson's R:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling:

Scaling is a technique to standardize the features present in the dataset in a fixed range.

Why Scaling?

In the dataset variables might have different range of values based on the type of data, where some variables have very low range and some variables have very high range values. So, it is extremely important to rescale the variables so that they have a comparable scale.

If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation.

Difference between Normalized Scaling & Standardized Scaling:

Normalization: (Xi-Xmin) / (Xmax-Xmin)

Standardization: (Xi-mean(x))/standard deviation(x)

Normalization	Standardization
Minimum and maximum values are used for scaling	Mean and standard deviation is used for scaling
Values are distributed between [0,1]	Values are distributed around 0 and no boundary range
Has better interpretation	Less interpretation
Removes outliers	Will not removes outliers
Used when we don’t know the distribution	Used when feature distribution is Normal or Gaussian

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

Let’s look into the formula of the VIF to understand this.

VIF = 1 / (1 - R-square)

- We can see that VIF is dependent on R-square/ Coefficient factor of the feature and its values are in range between [1 to infinite]
- It is possible to get the VIF as infinite when the R-square value is equals to 1 (i.e.) the feature has absolutely correlated to the field

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

Quantile-Quantile (Q-Q) plot, it is a plot of the quantiles of the first data set against the quantiles of the second data set.

The Q-Q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution

Use and importance of Q-Q plot:

- Sample size do not need to be equal

- Many distributional aspects can be simultaneously tested
- Explains whether two data sets come from populations with a common distribution
- Explains whether two data sets have common location and scale
- Explains whether two data sets have similar distributional shapes
- Explains whether two data sets have similar tail behaviour