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: As seen from the train set and test set, there are few independent variables that effects our Linear model. **Weekday**- People like to rent bikes to go to either offices or colleges. **summer, fall and winter**- People like to rent bikes in these season as it hasn't started raining or snowing Thus, instead of hiring a taxi or taking your private Car (which could be expensive considering Petrol and Diesel charges) they rent bikes. As **year** passes on, more people have attracted to rent bikes (as it may have become a trend). Hence, we should start some sound of Marketing or Discount rates to advertise more of our Bikes.

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

Answer: A dummy variable is a variable that takes on the values 1 and 0; 1 means something is true (such as age < 25, sex is male, or in the category "very much"). Dummy variables are also called indicator variables. It's important to drop_first=True during dummy variable creation, because the third (or the last variable as in our case- 4th) dummy can be explained as the linear combination of the first few. We always remove one level of your categorical feature which becomes the reference group during dummy encoding for regression and is redundant. Ex. Instead of having values for Gender- Male and Female, we could have a just one dummy variable which could contain values 0/1, where 1 could mean Female and 0 mean Male or vice versa. It reduces the code complexity and best expresses the relationship to your dependent variable.

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

Answer: Based on the pair plot, we have highest correlation for atemp, temp with the target variable

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

Answer: Following are few assumptions-
Assumption 1: We plot a linear graph b/w target variable and input variable to show they're linearly dependent.
Assumption 2: Plot a histogram to show that Error terms are normally distributed.
Assumption 3: Error terms have no pattern and are independent of each other→ by plotting a scatter plot.
Assumption 4: Plotting a scatter plot to show that the error terms have a constant variance (Homoscedastic) and thus the variance doesn't increase or decrease.

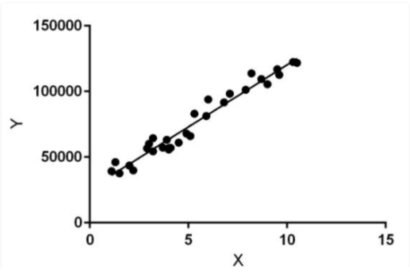
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Based on final model, we have year, fall and summer.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.
Hypothesis function for Linear Regression:

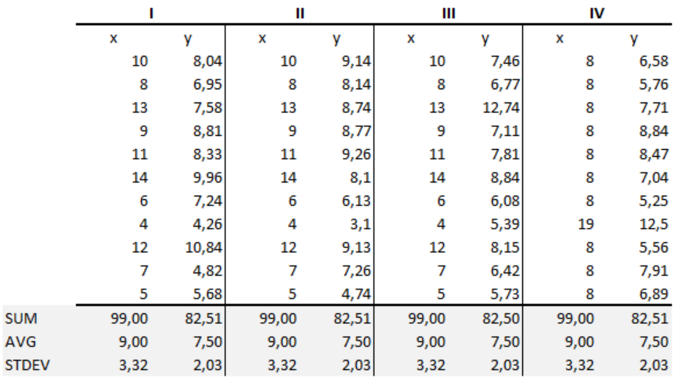
$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe's quartet in detail.

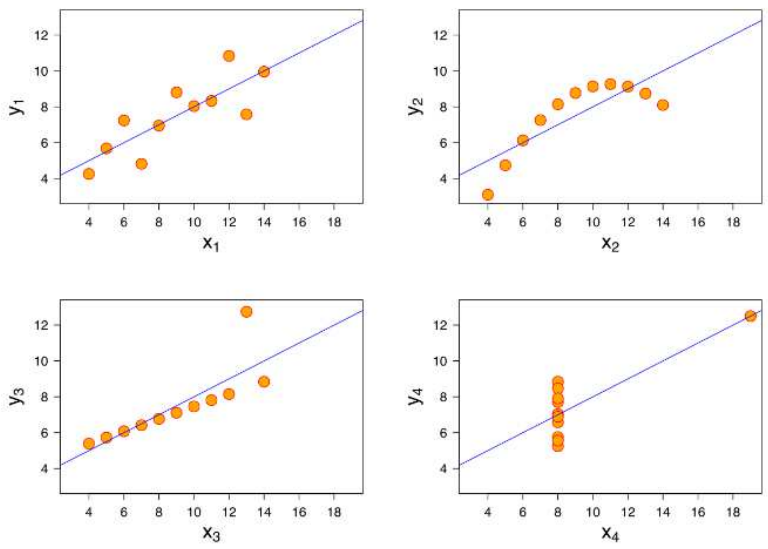
Answer: Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed.
Each graph tells a different story irrespective of their similar summary statistics.

(P. T. O)



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :
Mean of x is 9 and mean of y is 7.50 for each dataset.
Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

3. What is Pearson’s R?

Answer: The Pearson correlation coefficient, also called Pearson’s R, is a statistical calculation of the strength of two variables’ relationships. In other words, it’s a measurement of how dependent two variables are on one another. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Pearson’s R depicts the extent that a change in one variable affects another variable. This relationship is measured by calculating the slope of the variables’ linear regression.

The value of Pearson’s R can only take values ranging from +1 to -1 (both values inclusive). If the value of r is zero, there is no correlation between the variables.

If the value of r is greater than zero, there is a positive or direct correlation between the variables. Thus, a decrease in first variable will result in a decrease in the second variable.

If the value of r is less than zero, there is a negative or inverse correlation. Thus, a decrease in the first variable will result in an increase in the second variable.

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

Answer: Scaling means adjusting data that has different scales so as to avoid biases from big outliers. Feature scaling is a method used to standardize the range of independent variables or features of data.
Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
Formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
Formula:

$$X' = \frac{X - \mu}{\sigma}$$

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

Answer: If two Xs are perfectly correlated
VIF = 1/(1-1)= 1/0 = infinity

The VIF is efficiently calculated (not by running a series of regressions) but as the diagonal element of the inverse of the correlation matrix of the predictors. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Answer: In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples.