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?**

Ans:

As per the Data Visualization for categorical columns using the boxplot we can infer the following points from various features of Data

- Trends of bookings seen majorly in the month of may, june, july, aug, sep and oct. Trend seems to be starting of the year till SEP-OCT and there after seen gradual decrease in demands as moving to end of an year
- Clear weather season attracted more booking which seems obvious.
- Considering the box plot Fall season mostly in September, October, November there are more bookings in as moving on from 2018 to 2019 count of bookings are gradually increased as per Data observations.
- WED, THUR,FRI.SAT are mostly seems to attract customers with number of bookings compared to the other days of the week.
- When there is working Day booking are seen good in numbers which seems opposite as on holidays as people mostly like to be at home and use the bikes only for working time.
- 2019 seems to have good number of booking compared to 2018, which shows good exposure to business and growth in overall acceptance of business model.
- 2018 to 2019 every month we have seen increase in number of bookings majorly for working Days compared to holidays.

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

Ans:

drop_first – An optional component set to 'False' by default and is set to 'True' if the first level from the input categorical data is to be removed while converting to dummy variables.

It helps in reducing the extra column created during dummy variable creation which reduces the correlations created among dummy variables.

Syntax -

drop_first: bool, default False, which implies whether to get a-1 dummies out of a categorical levels by removing the first level.

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

Ans:

'atemp' variable has the highest correlation with the target variable.

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

Ans:

There are 5 basic assumptions of Linear Regression Algorithm These assumptions are just a formal check to ensure that the linear model we build gives us the best possible results for a given data set and these assumptions if not satisfied does not stop us from building a Linear regression model.

→**Linear Relationship between the features and target:** Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

→**No Multicollinearity between the features:** Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.

→**Homoscedasticity Assumption:** Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all

values of the independent variables. A scatter plot of residual values vs predicted values is a Goodway to check for homoscedasticity

→**Normal distribution of error terms:** Normal distribution of the residuals can be validated by plotting a q-q plot.

→**Little or No autocorrelation in the residuals:** Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.

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

Ans:

As per our final Model, the top predictor variables that influences the bike booking are:

1. Year (yr) - A coefficient value of '0.2445' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2445 units
2. spring - A coefficient value of '-0.2726' indicated that, w.r.t spring, a unit increase in spring variable decreases the bike hire numbers by 0.2726 units.
3. Light_snowrain- A coefficient value of '-0.3245' indicated that, w.r.t Light_snowrain, a unit increase in Light_snowrain variable decreases the bike hire numbers by 0.3245 units

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. it performs a regression task. Regression models predict a dependent (target) 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 the dependent and independent variables, they are considering, and the number of independent variables being used.

Mathematically the relationship can be represented with the help of following equation −

Y = mX + c Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model −

✓ **Multi-collinearity –** Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

✓ **Auto-correlation –** Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

✓ **Relationship between variables – Linear** regression model assumes that the relationship between response and feature variables must be linear.

✓ **Normality of error terms** – Error terms should be normally distributed

✓ **Homoscedasticity – There** should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

Ans:

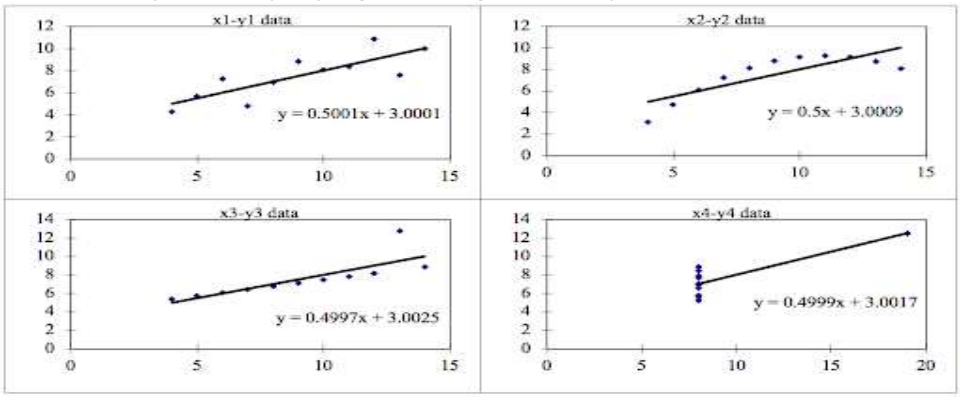
Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions, so they look completely different from one another when you visualize the data on scatter plots. Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



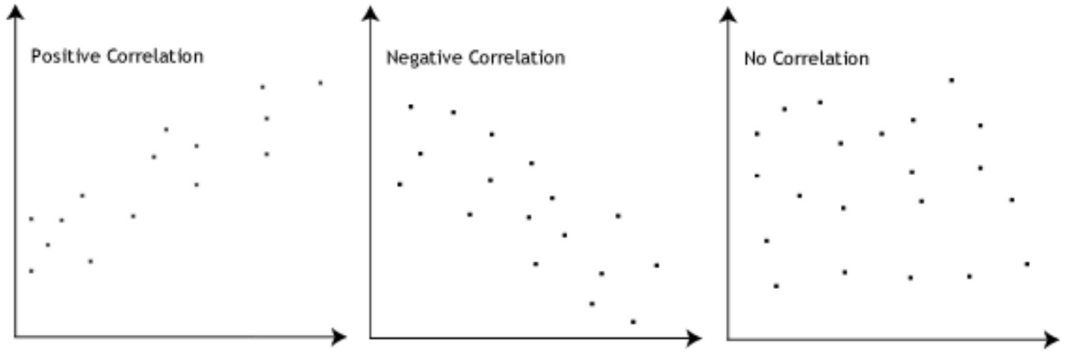
We can describe the four data sets as: ANSCOMBE'S QUARTET FOUR DATASETS
Data Set 1: fits the linear regression model well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

Ans:

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Ans:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

.NO.	Normalization	Standardization
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans:
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans:
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour