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?

Different levels of each categorical variables have certain impact on the dependent variable. To elaborate – the demand of shared bikes depends on whether it is Spring season or not, whether it is a working day or not, the season, which month of the year it is etc.

From Exploratory Analysis of Data, we observed the following-

a. There is higher demand for shared-bikes in Summer, Fall and Winter and people shrink back to use bikes in Spring season. The median of Spring is considerably lower than the other three seasons.
b. The demand for shared bikes services has considerably increased from year 2018 to 2019, showing change in people's inclination towards shared-bikes usage as mode of transportation.
c. For each year, the period between April thru Oct has a higher demand for shared bikes which is in-lines with the seasonal trend.
d. Over the holidays, the number of users vary more drastically and the median is lower whereas, for non-holidays the user count is concentrated between range 2500 to 6000 and median is higher as well. This could be because more users opt for shared-bike to reach their workplace. Similar observation can be made for working day.
e. Over the week, the users prefer shared bikes on Thursday, Friday and Saturday more while the demand reduces on Sundays.

Similarly, we see in the Multiple Regression Model, the features like Year, Summer, Winter season, working day, Saturdays and September are levels of different categorical variables which are important for determining the demand for shared bikes (independent variable). All these variables have a positive coefficient which mean that an increase in these variables (individually or combination) will lead to increase in the demand of the shared bikes.

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

As each level of categorical variable is encoded using binary (0 or 1), the level which has all the dummies represented by 0s becomes redundant and need not be shown separately. This level can essentially become base level and other dummy variables are then represented on top of this base level. By dropping the first column, we reduce the number of dummy variables by one, and ensure that each variable represents the difference from the base level. Thus, for n-level of categorical variable, n-1 dummy variables are sufficient to represent each of the levels of the categorical variable. In turn, we have one less column or feature to consider in model building which also reduces the chances of multicollinearity or correlation among the dummy variables.

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

The numerical variable 'atemp' has the highest correlation with the target variable of about 0.63. This is after removing the columns 'registered' and 'casual' (removed based on details provided in

problem statement of the assignment; 'registered' has highest correlation with the target variable otherwise.)

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

Assumptions of Linear Regression in model building and their validation was done in following manner -

- Linear relationship – Each independent variable must be linearly related to the independent variable. We have plotted graph to check the relation between each numerical variable with the target variable.
- Multicollinearity – The independent variables should not be highly correlated with the target variables. We have checked the Variance Inflation Factor (VIF) for features selected and taking a threshold of 5 we have dropped the variables which have high VIF i.e. the ones that are highly collinear to other variables.
- Multivariate normality – The residual or the error terms (difference between the observed and predicted values of target variable) should be normally distributed. We have plotted a graph to check the distribution of the error terms.
- Overfitting – As the features are added, the model can end up memorizing the training data set and will fail to generalise, this is visible if the training accuracy is high while the test accuracy is very low. We have dropped several features which we believed were not suitable for the model. In our case, the training test accuracy is 0.79 and test set accuracy is 0.75, the difference is not too wide.
- Independence of error terms – The error terms should not be dependent on one another. Again, we have plotted a scatterplot to check the dependency of the error terms.
- Homoscedasticity – The residual or the error terms have constant variance.

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

Based on the final model, the top 3 features contributing significantly towards the demand of shared bikes are temp, windspeed and year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a technique of finding out how two or more variables are related by using a straight line. Linear Regression assumes a linear relationship exists between the dependent and independent variable(s). It can help to predict a target or dependent variable based on one or more independent variables. Interpolation is possible however extrapolation is not possible with linear regression technique. We determine a best-fit line that minimizes the difference between the actual and predicted values of the outcome variable.

Linear Regression Algorithm is an essentially a supervised learning, i.e., it uses labeled data (with known outcome values) to train a model that can make predictions for new data. In Simple Linear

Regression, there is one independent variable and one dependent variable. In case of Multiple Linear Regression, there are 2 or more independent variables and one dependent variables.

Steps of Linear Regression Algorithm:

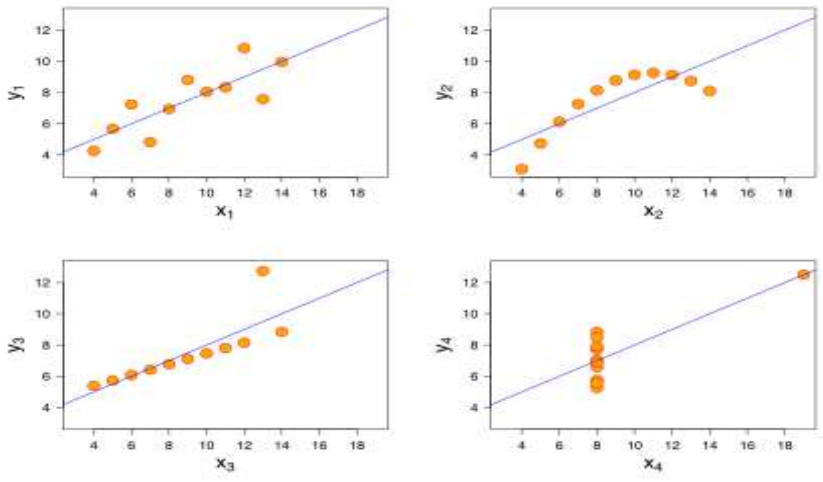
1. Reading, Understanding and Visualizing the Data – First we upload the dataset, then we perform descriptive statistical methods to understand the data. We perform sanity check for missing values etc. We visualize the data to check various aspects like if any linear relationship exists between target and predictor variables using pair plot, correlations among different variables, analyse categorical variables their count/frequency etc.
2. Prepare the data – This step involves, understanding the types of variables whether continuous or categorical. Variables with text values/levels are converted to binary or suitable encoding to numbers. 'n' Categorical variables are converted to 'n-1' dummy variables.
3. Splitting the Data into Training and Test Data Set and Rescaling – The dataset is split into Train and Test set Mostly (70-30 or 80-20 split). The training data set is used to train the model while the test set should remain un-seen to the training model. We also rescale the values of variables of training set once the data is split. Rescaling is done to adjust the values of the dataset on a comparable or common scale.
4. Building the Model – We can build a model by adding variables one-by-one, wherein we take the feature which we feel is most important as starting point and we keep adding more variables and checking the performance of the model as we progress.
   Other method is to build model with all the variables and then drop the irrelevant or insignificant variable one-by-one.
   We can also start with an automated approach where the pre-mentioned number of features are selected by the system using Recursive Feature Elimination technique after that we drop the variables one-by-one from the selected features.
   Depending on which library we are using for model building, we may or may not need to add a constant to the model.
   The training data is fit to the model using OLS method. Each time we fit a model and the coefficients of the features are determined. We then check on certain criteria – R-squared, p-values of features and VIF to determine how well the model is working, the significance of the features and correlation among the features. This is required to determine which feature to eliminate from the next updated model to make like less complex. It can several iterations on this process to arrive at a final model which does not compromise on the threshold levels of p-values and VIF.
5. Residual Analysis of Train Data - Once we are satisfied with the model, we do a Residual analysis of the train data. Here, we check how good the model holds on the assumption of Linear Regression model viz.
   a. Residual or Error terms are normally distributed.
   b. Independence of Error terms with target or predictor values.
   c. Homoscedasticity i.e., residual or the error terms have constant variance.
6. Making Prediction using Final Model – We now use the model to predict the values of the test set data. For this we need to ensure to divide test data to target and predictor variables, add constant, rescale, select and drop the features that are irrelevant.
7. Model Evaluation – We evaluate the model by comparing the test set data target variable vs the predicted target variable using model. We also check R-squared value for test data set to ensure that the accuracy of model does not varies too much from the train data set model. Otherwise, it would mean, overfitting.

2. Explain the Anscombe's quartet in detail.

Consider the datasets provided below. The four data sets may look similar when we consider the simple descriptive statistics like mean, standard deviation however when we plot the data points, we observe that each dataset may have a different distribution. Thus, it is very important to visualize the data before applying any algorithms to build models. This concept is known as Anscombe's quarte. It helps to identify any anomalies in the data (ex. Outliers, etc.). The importance of Anscombe's quarte is obvious for example - when we want to check whether the Linear Regression model can be built for certain data. We first have to visualize the data to check if there is any linear relationship between the dependent and the independent variables. In the example below the third dataset can fit the linear regression model pretty well, while the others do not have linear relationship.

(source for dataset and plots: Anscombe's quartet - Wikipedia)

Anscombe's quartet							
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
Mean = 9	**7.50090**	**9**	**7.50090**	**9**	**7.5**	**9**	**7.50090**
Std Dev. = 3.3166	**2.0315**	**3.3166**	**2.0316**	**3.3166**	**2.0304**	**3.3166**	**2.0305**



3. What is Pearson's R?

In simple words, Pearson's R gives the strength of association of two variables i.e. how one variables changes when the other variable is changed. It is used to measure the linear correlation between the two variables. It ranges from -1 to 1; Pearson's r of Negative 1 implies the two variables are change in opposite directions perfectly; a Positive 1 implies the two variables move in same direction again perfectly and a Pearson's r of 0 implies that there is no correlation between the two variables or no linear relationship.

The Pearson's correlation coefficient formula is

r = [n(Σxy) − ΣxΣy]/Square root ofV[n(Σx2) − (Σx)2][n(Σy2) − (Σy)2]

An important caveat in using Pearson's r is that it does not indicate any cause-and-effect relationship. It only talks about how or to what extent the variables are associated but it does not show which variables causes the change (independent variable) and which one is the dependent variable. The same is clear from the formula above, as either of X or Y can be dependent or independent variable the r value will not change based on this.

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

Usually, the values of various variables in the dataset have different magnitude, some may be too high (ex. in thousands or millions) while others may be too small (ex. Decimals or single digit). Scaling is a data pre-processing technique wherein the values of different variables/features in a dataset are re-scaled to a common scale so that they are comparable. It is an important step because 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.

It is recommended to use standardization or normalization scaling method so that the units of the coefficients obtained are all on the same scale.

Normalization (Min-Max scaling) – This method is used to adjust the values of features to common scale ranging between 0 and 1. It is also known as Min-Max Scaling as it utilizes the Minimum and Maximum values of the feature to rescale the value.

$$X' = \{X - Xmin\}/(Xmax - Xmin)$$

When the value of X is minimum, the numerator becomes zero and hence X' is 0; when the values of X is maximum, the numerator is equal to denominator and hence X' is 1.

It retains the shape of original distribution of the values.

Standardisation (mean-0, sigma-1) – In this technique the values of features are adjusted such that they are centered around the mean with standard deviation of 1. That is, the mean of the feature becomes zero and the distribution has one standard deviation.

The formula for Standardization is given by –

$$X' = (X - \mu)/\sigma$$

where μ is mean and σ is standard deviation.

There is no specific range of the rescaled values in case of standardisation.

It modifies the shape of original distribution of the values.

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

The Variance Inflation Factor (VIF) is given by –

$$VIF = 1 / (1 - Ri^2)$$

Where Ri^2 is R-squared of the i-th variable (the coefficient of determination for regressing the i-th independent variable on the remaining ones).

Thus, when the R-squared is high or approaches 1, the denominator will be closer to zero hence the VIF will become infinity.

In terms of the features or variables of dataset, VIF of infinity means that the variance of the remaining independent variables can be clearly predicted from the i-th independent variable indicating they are highly corelated and multicollinearity exists.

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

A quantile-quantile plot or the Q-Q plot is used to visualize and compare probability distribution of two datasets in order to check whether or not a dataset fits a certain theoretical distribution. Basically, it is a scatter plot that shows the relationship between the quantiles of two datasets. So the quantiles of each datasets are along X and Y-axis respectively.

Since the Q-Q plot is used for comparison of two datasets, it is intuitive to conclude that when the plot is straight line, the two datasets are same or the quantiles are taken from same datasets. On the other hand, if there is deviation, then we can check the relationship between the two quantiles distribution whether they follow any other theoretical distribution like normal distribution or log-normal distribution. Thus, it is the deviation from the straight line that can provide useful insights in to the relationship/differences between the two given datasets.