**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:***

1. Why is it important to use drop\_first=True during dummy variable creation?

***Answer:***

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

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

***Answer:***

*correlation between temp and atemp is 0.99 which is almost 1.*

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

***Answer:***

*Based on VIF and p-value we will find the best fitted line.*

*Best fitted line on our analysis is*

*cnt= 0.246 X yr - 0.083 X holiday - 0.198 X Spring - 0.321 X Light rain\_Light snow\_Thunderstorm - 0.090 X Mist\_Cloudy +0.063 X 3 + 0.123 X 5 +0.148 X 6 +0.153 X 8 + 0.193 X 9 -0.049 X Sunday + 0.126 X 7 + 0.116 X 10*

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

***Answer:***

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

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

***Answer:***

*Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range rather than trying to classify them into categories .*

*There are two main types:*

* *Simple regression :*

*y=mx+b*

*x represents our input data and y represents our prediction.*

* *Multivariable regression :*

*f(x,y,z)=w1x+w2y+w3z*

*The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation*

1. Explain the Anscombe’s quartet in detail.

***Answer:***

*Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.*

*Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.*

* *Dataset 1: this fits the linear regression model pretty well.*
* *Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.*
* *Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model*
* *Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model*

1. What is Pearson’s R?

***Answer:***

*Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.*

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

***Answer:***

*It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.*

*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 have to do scaling to bring all the variables to the same level of magnitude.*

|  |  |
| --- | --- |
| ***Normalized scaling*** | ***Standardized scaling*** |
| *Min-Max Scaling* | *Standardization Scaling* |
| *It brings all of the data in the range of 0 and 1.* | *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).* |
| *sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.* | *sklearn.preprocessing.scale helps to implement standardization in python.* |

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

***Answer:***

*An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.*

1. 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, 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.*

* *It can be used with sample sizes also.*
* *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.*