Assignment Answer :-

1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable? (3 marks)

Answer

- Temperature has high positively correlation and when there is more temp* the number of bike hire are increase.
- Number of bikes hired is very less in season spring then other.
- Numbers of bike hires are very high from May-to-October.
- Number of hired bikes quantity are very less in weather situation-3.

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

Answer

is important to use, to helps in reducing the extra column created during dummy variable creation.

2. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Answer

* Temp and atemp have highest correlation with "cnt".

3. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Answer

The assumption of Liner Regression is validated with Residual Analysis after building the model. To verify the normality of the error, we can use dist-plot to see residual value is distributed normally. If the

curve is skewed then it show that the model is biased.

4. Based on the final model, which are the top 3 features contributing significantly towards
explaining the demand of the shared bikes? (2 marks)

Answer

Temp : Temperature with highest value of coefficient of '0.564492' indicates the a unit increase in temperature will increse the bike hires by 0.564492 units.

Yr : Year with second highest coefficient value indicate each year to year potential grow in the number of bike hire by '0.231122' units

weathersit_3: Weather situation 3 indicates the negative correlation means with a unit increase in weather_sit3 there is a decrease in the bike hire by 0.315252 units

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer

Data Understanding:

- understand each feature and the importance and how they can impact the target variable. Which are independent variable and dependent variable so on

Data Preparation and Data Cleaning:

- Check for missing values and junk values, if any columns have more than 50% missing values then drop those columns.
  And fill the missing values with mean/mode
- Assess which are important and which are unwanted feature.
  Remove all unwanted columns from the dataset

Creating Dummy Variable:

- I identify the categorical variables, and convert them into numerical representation by crating dummy variables and It is import to drop first column because it reduces the number columns for building correlation among dummy variables. The first column can be derived from

the remaining variable and it has no added value, hence it is better to remove it.
- Always if N levels of predictor variable are there then we need n-1 dummy variables
-

EDA Process and data visualisation:
- Understand the impact of each and every feature how they change the target variable.. data visualisation will help to understand correlation between variables. Plotting pair-plot and boxplot and correlation matrix we can understand the relation.

Rescaling:
- the collected data will be in differentlevels, units and magnitude. If the scaling is not done one may end up build incorrect modelling because different units and mangitudes. Since the numeric variable are represented in different ranges it will be hardto work on larger numbers so we need to first rescale the value to lesser range for building easy and fast model
- We have two rescaling methods
  Minmaxscaling
  Standardization scaling

Building the model:
- split dataset into the train and test data
- add constant
- create linear model using Ordinary Least Square
- fit the model

rebuild the model:
- when there are Multi-collinearity we need to usedifferent approaches to build the model
  - we can build the model with all the variable
  - We can build model by adding variable one by one and check R^2 value to check significance
  - Use RFE method to improve R^2 value by eliminating the feature which has high P value and High **VIF** (which indicate. Significance and Variance)

Model assessment and Prediction

- using Residual analysis a check error terms are normally distributed or not by plotting the dist-plot
- Predict the value of target variable using the final model

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer

A group of four data sets which are nearly identical in statistical property but they look different on graphs.
Anscombe's quartet shows the importance of plotting data to confirm the validity of the model fit.

What is Pearson’s R? (3 marks)

Answer

It is a measure of linear correlation between two sets of data. It is the ratio between thecovariance of the two variables and the product of their standard deviations. It is normalised measurement of the covariance which is usually between -1 to -1.

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)

Answer

- Usually the collected data will be in different levels, units and magnitude. If the scaling is not done one may end up build incorrect modelling because different units and magnitudes. It will be so crusial to work on larger numbers of row & columns so we need to first rescale the value to lesser range for building easy & fast model
- We will have two rescaling methods

  MinMaxScaling:

  It brings all the data in the range of 0 to 1

formula: **X = (x – x_min) / (x_max – x_min)**

Standardization scaling

It brings all the data into a standard normal distribution.

Standardisation: x = (x – mean(x))/std(x)

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

Answer

VIF is basically indicator of variance, mean how the variable is correlated.

If VIF is 'infinite' which indicate the perfect correlation between the two independent variables.

When R^2 becomes 1 which means there is no difference in observed data and the fitted values.

VIF = 1/(1 – R^2) becomes infinity. We need to perfect multicollinearity by removing the variable from the dataset.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer

It is to find if two sets of data come fromthe same distribution.
Q-Q plots is plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction of points. Below the given value that is 0.3 (or 30%) quantile Is the point at which 30% of the data fall below and 70% fall above the value
From Q-Q plots we can find out if two data sets:

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour