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? (3 marks)

Answer: Listed below few inferences on effect of categorical variable on dependent variable.

- Bike demand is zero for heavy snowfall. Light snowfall also shows very low demand of bikes.
- Clear weather is preferred by riders and show spike in demand. Cloudy/misty weather also show some demands of bikes.
- No variation observed in weekday demand of bikes as all days show almost same demand.
- On holidays less demand compare to non-holidays.
- Fall has highest demand, followed by summer and winter and least demand is seen in spring may be due to snow. This is validated by the above weather conditions as weather is clear during fall and summer seasons.
- May to October show spike in bike demand. January had least demand for bikes.

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

Answer: Since we can predict the dropped variable by existing dummy variables which will allow us deliver same information with least number of variables. It will allow us to avoid redundancy in data. It provides ease of interpretation between base state and effect of state.

Example: If we have gender column with value as male and female and we created dummy variables for this column as male and female. In male column if its female it will be 0 else 1. In female column if its male value will be 0 else 1. We don't need both these columns to identify gender. If we drop female we can figure out if its male or not using female column when value is 0.

The general rule in linear regression for dummy variables is have n-1 columns for n variables/levels.

Python pandas allows this by property drop_first while creating dummy variables by dropping first variable.

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

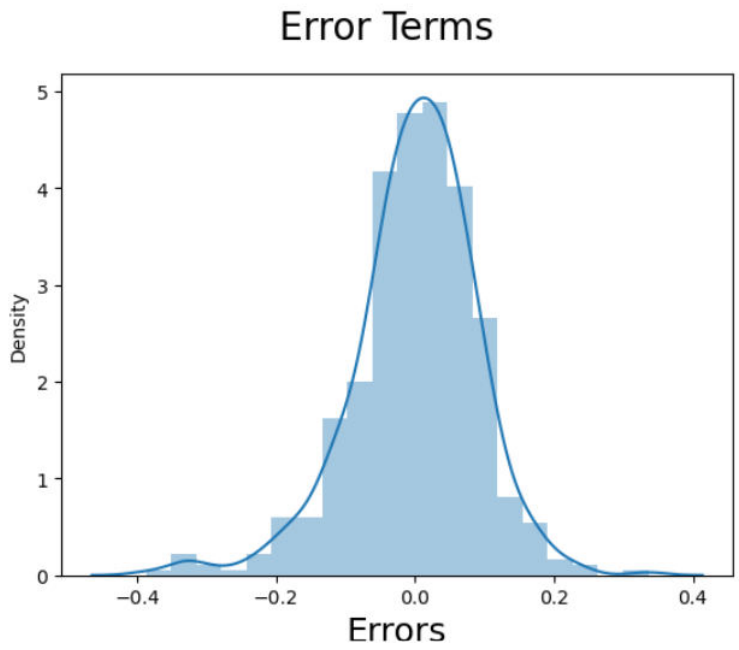
Answer: temp and atemp has highest corelation with target variable(0.63).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: We did residual analysis in which we calculated error terms using y train and y train predicted. Residuals should show normal distribution and mean as 0. As per below screen shot, we saw normal distribution for residuals and mean also as 0

```
In [760]: # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((y_train - y_train_pred), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)
```

```
Out[760]: Text(0.5, 0, 'Errors')
```



We also saw linear relationship between dependent and independent variables.

We verified no multicollinearity on final model with correlation matrix. All variables have VIF < 5 in final train data model.

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

Answer: Below are the top 3 features contributing significantly on demand of bikes

- Temp: Temperature has 0.5706 correlation coefficient. Unit increase in temp will increase demand of bikes by 0.5706 units.
- Light_snowrain: light snowrain has -0.2367 correlation coefficient. Unit increase in light snowrain will decrease demand of bikes by 0.2367 units.
- Yr : Year has 0.2289 corelation coefficient. Unit increase in year will year demand of bikes by 0.2289 units.

```
Out[838]: const            0.225646
          yr               0.228914
          holiday         -0.097964
          temp             0.570606
          hum             -0.173973
          windspeed       -0.186706
          Summer           0.089525
          Winter           0.140200
          Sep              0.106731
          Light_snowrain  -0.236675
          Misty           -0.051831
          dtype: float64
```

General Subjective Questions

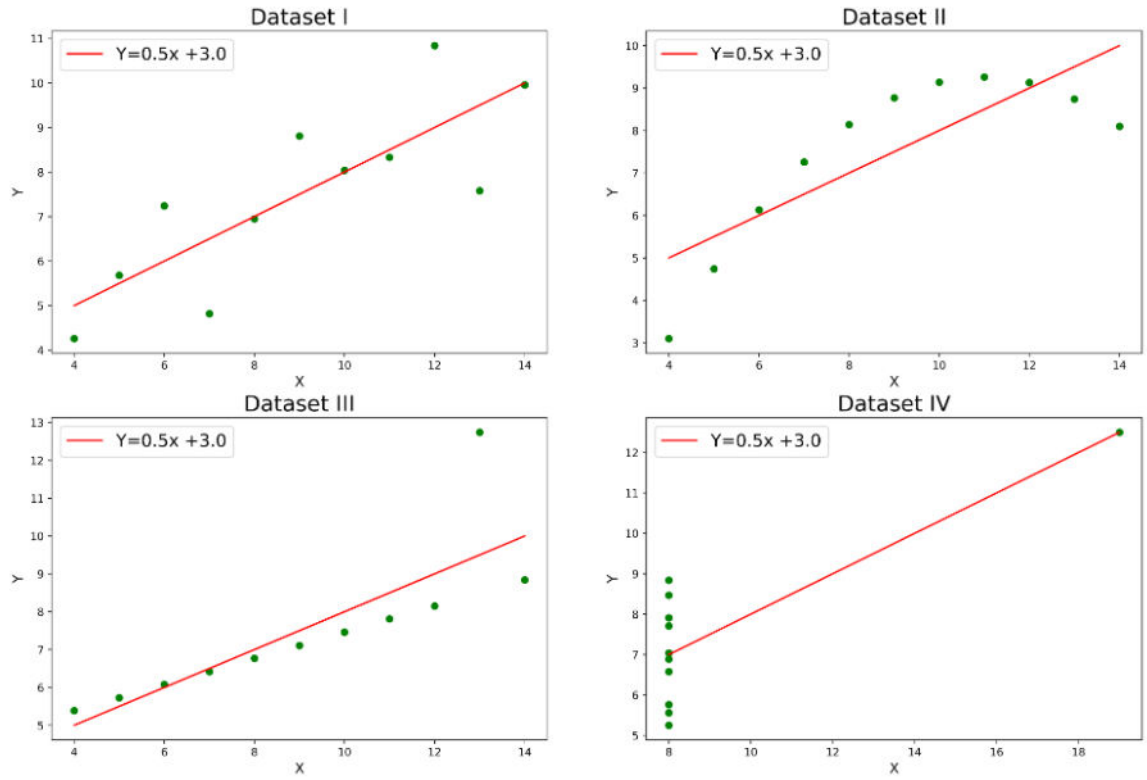
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

- Understanding the data
- Take care of missing data if any
- By looking at data and its data dictionary drop unwanted columns
- Do data mapping wherever applicable.
- Do extensive EDA (univariate, bivariate and multivariate analysis)
- Data preparation for model. Create dummy variables for categorical variables
- Split the data in train – test datasets
- Rescale numeric variables in training dataset
- Divide training data in X and Y
- Select columns for model if many available
- Build a model.
- Drop columns having high VIF and P-value
- Rebuild the model
- Predict Y on training dataset
- Perform residual analysis on final model by calculating error terms and check normal distribution.
- Prepare test data by rescaling numeric variables and selecting only columns available in training data in in significant model.
- Divide test data in X and Y. Calculating r squared and adjusted r squared on test data.
- Visualize fit on the test data

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

Answer: Anscombe's quartet has 4 datasets with identical descriptive statistics properties like mean, variance, correlation, r squared and linear regression etc. but very different visual distribution when plotted scatter plot of these datasets. It demonstrated importance of graphical representation and not just replying on statistical properties and how outlier's impact on these properties and graphs.



All above datasets have same correlation coefficient.

The first dataset shows some linear relationship.

The second dataset shows nonlinear relationship.

The third shows perfect linear relationship.

The fourth show how single outlier made correlation coefficient higher.

It shows how exploratory data analysis plays significant role while analysing data rather than solely relying on the stats data.

3. What is Pearson's R? (3 marks)

Answer: Pearson's R(r) is a way of measuring linear correlation.

It is number between -1 and 1 showing negative positive relation between two variables.

If the coefficient is between 0 and 1 then relationship between two variables is positive that means, if one variable increases other increases and vice a verse, also called as moving in same direction. Coefficient value towards 1 is indication strong positive relationship, however towards 0 its weak positive relationship.

If the coefficient is between 0 and -1 then relationship between two variables is negative that means, if one variable increases other decreases and vice a versa, also called moving in opposite direction. Coefficient value towards -1 is indicating strong negative relationship, however towards 0 its weak negative relationship.

If coefficient is 0 then there is no relationship between 2 variables.

Its used quantitative variables have linear relationship and data don't have outliers and normally distributed.

Formula:

R=sum(Xi-x')(Yi-y')/square root(sum((Xi-x')^2*(Yi-y')^2))

R=pearson's correlation coefficient

Xi=values of x variable in sample

x'=mean value of x variable

Yi=values of y variable in sample

y'=mean values of y variable

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

Answer: It is data preparation step while building linear regression model in which values of variables will be normalized in particular range as all the variable might have different measurement scale.

Since different variables in data have different units and range. Building model on this data will not take unit in consideration and will be flawed. Scaling will help to bring all variables in same range to avoid above stated problem.

Scaling will only impact the correlation coefficient and not any other statistical properties.

Normalised Scaling: It brings all data in range of 0 and 1. Also, knows as min-max scaling.

Min – Max scaling=x=min(x)/max(x)-min(x)

X is the variable we are scaling in given dataset.

Standardised scaling: It replaces the variable values with there z score which will in the end bring all data in standard normal distribution with 0 mean and standard deviation as 1.

Standardised x=x-mean(x)/standard deviation(x)

X is the variable we are scaling.

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

Answer: It indicated the perfect correlation between the variables.

Its is calculated by formula, VIFi=1/1-R^2i

I indicated ith variable.

When R square of ith variable is one VIF goes infinite showing perfect relationship of independent variable.

It happens due to multicollinearity between the variables. If there are duplicates or the data is redundant in the data it impacts VIF. Also, if the variables are not scaled in same units it does have impact on VIF.

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

Ans: It is a quantile quantile scatter plot. It's created by plotting 2 different quantiles against each other, on if which is on actual distribution against which hypothesis will be tested and other is that of variable which we are testing hypothesis for.

Useful for understanding test data distribution of variable against train data distribution.

Can show any deviations in the test data, could be result of skewed or outlier data in variables.