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 : From analysing the categorical variables from the dataset provided, the following were the points that could be drawn about their effects on the dependent variable , cnt

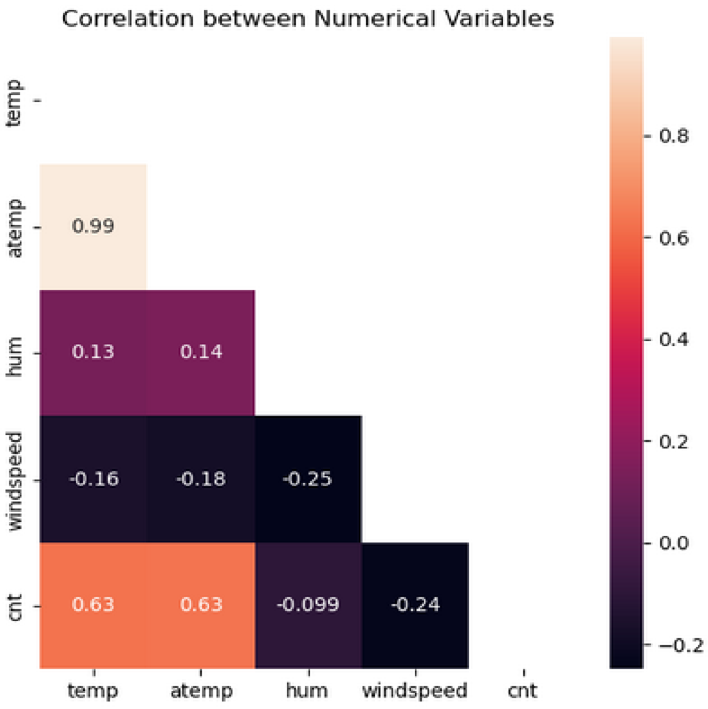
1. When comparing season with cnt, it was found that fall has the highest median, so it might be the most optimal weather condition for customers to ride bikes followed by Summer
2. On comparing the months (mnth) with count (cnt) , it reflects the season as fall months seems to have a higher median here as well
3. Comparing weekday, there seems almost same median, how ever the spread on saturday and wednesday seems to be very high.
4. On comparing the weather, it's is seen that Clear weather is most optimal for bike renting and it's the least optimal when it's snowing or raining. Hence sales will be higher on clear days and significantly less when it's snowing/raining
5. Comparing Holiday, People seems to rent more on non-holidays possibly due to a large number of customers using the service for daily commute. And during holidays, they might travel as a family or group and might prefer cars or personal vehicles.
6. Working day and non working day jabe a similar median bit spread is higher for non working days, it could possibly be that for non working days , people might not need to travel via bike
7. Median of Bike renting is increasing with year, 2019 clearly has a higher median compared to the year before. It could be that people are getting more interested in bikes as a mean of transport due to popular trends and getting interested in going green

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

Ans : the drop_first parameter is used to prevent multicollinearity issues in regression models.
drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.
drop_first allows you whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels)
So when you set drop_first = True, then it will drop the reference column after encoding.

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

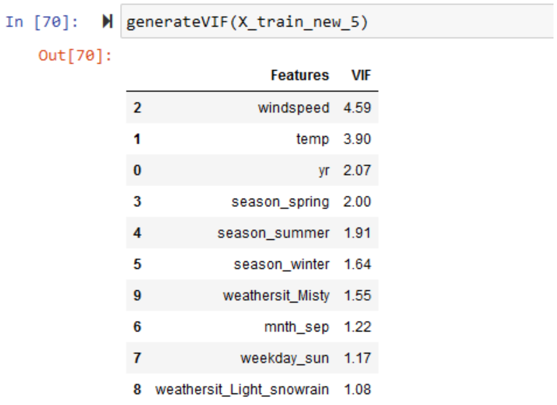
Ans : In the Pair-Plot showing the correlation between Numerical Variables, ‘temp’ has the highest correlation with the target variable, ‘cnt’ , with a value of 0.63 as can be seen from the screenshot below



Looks like temp and atemp has a hight correlation here

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

Ans : since validating the assumptions of linear regression is crucial to ensure the models assumptions are met and that model is appropriate for data, I first performed Residual Analysis., created a histogram of the residuals, If the residuals are approximately normally distributed , the assumptions of normality are met. Linear relationship validation using CCPR plot was done as well. Also checked for multicollinearity, Assess multicollinearity among predictor variables. Calculate the variance inflation factor (VIF) for each predictor. A high VIF indicates potential multicollinearity issues.



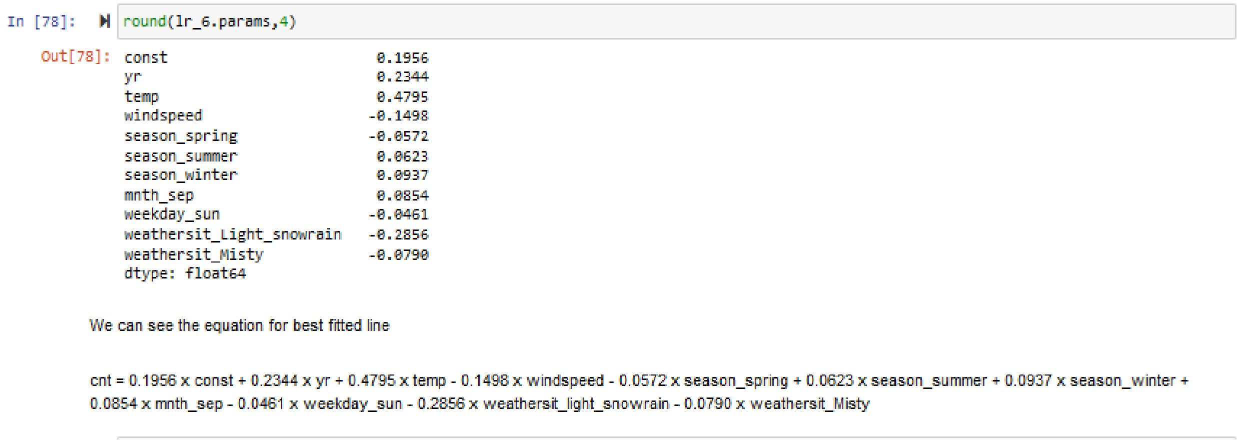
```
In [70]:  generateVIF(X_train_new_5)
  Out[70]:
```

	Features	VIF
2	windspeed	4.59
1	temp	3.90
0	yr	2.07
3	season_spring	2.00
4	season_summer	1.91
5	season_winter	1.64
9	weathersit_Misty	1.55
6	mnth_sep	1.22
7	weekday_sun	1.17
8	weathersit_Light_snowrain	1.08

The VIF was under 5 so it looked good

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

Ans :



```
In [78]:  round(lr_6.params,4)

Out[78]: const                        0.1956
         yr                           0.2344
         temp                         0.4795
         windspeed                   -0.1498
         season_spring               -0.0572
         season_summer                0.0623
         season_winter                0.0937
         mnth_sep                     0.0854
         weekday_sun                 -0.0461
         weathersit_Light_snowrain   -0.2856
         weathersit_Misty            -0.0790
         dtype: float64
```

We can see the equation for best fitted line

cnt = 0.1956 x const + 0.2344 x yr + 0.4795 x temp - 0.1498 x windspeed - 0.0572 x season_spring + 0.0623 x season_summer + 0.0937 x season_winter + 0.0854 x mnth_sep - 0.0461 x weekday_sun - 0.2856 x weathersit_light_snowrain - 0.0790 x weathersit_Misty

From the screenshot above , it is seen that yr , temp and weathersit_Light_showrain are the top 3 features contributing significantly towards explaining the demand of the shared bikes. They are listed below in order of highest contributor as :

1. temp = +0.4795
2. Yr = +0.2344
3. weathersit_Light_snowrain = -0.2856

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : Linear Regression is a statistical method used for modeling the relationship between a dependent variable. It assumes that the relationship between the variables can be approximated by a linear equation. The goal of linear regression is to find the best-fitting linear equation that can predict the dependent variable based on the independent variables
Linear regression seeks to find a linear equation of the form:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n$$

Where y is dependent variable that we are trying to predict
X1, x2 etc are the independent variables used to make the predections
B0, b1 etc are the corefficents

A linear relationship can be either positive or negative, Positive relationship is when both independent and dependent variable increases, and negative when the independent variable increases when the dependent variable decreases

For Linear regression we make some assumptions

1. MultiCollinearity : we assume that there is very little or no multicollinearity in the data
2. Autocorrelation : we assume there is very little autocorrelation in the data
3. Assume the relationship between the target variable and the feature variables must be linear
4. We assume that the error terms should be morally distributed
5. And that there should be no visible pattern in residual values

The following are the steps involved in Linear Regression :

1. Formulating the Linear Equation
2. Estimating coefficients/ Training
3. Fitting the model
4. Making predictions
5. Assumptions of linear regression (mentioned previously)
6. Evaluating the model using MSE or R2
7. Interpreting Coefficients

2. Explain the Anscombe's quartet in detail.

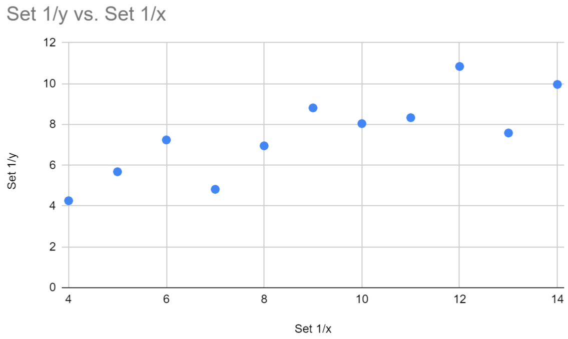
Ans : Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.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.

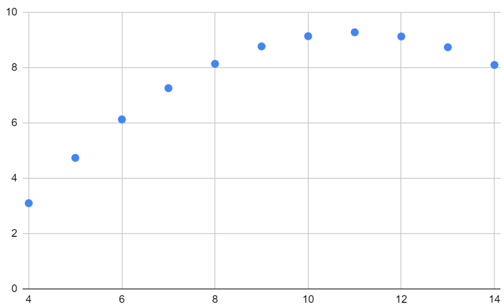
Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model.

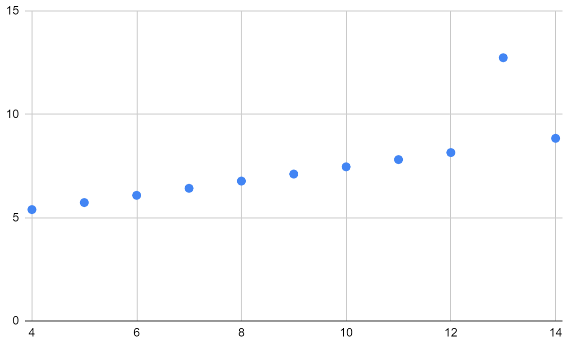
For Example :

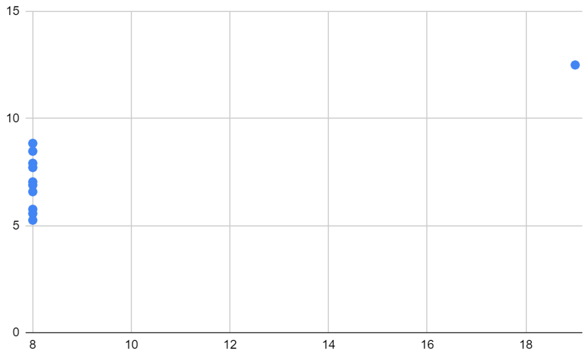
These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Set 1		Set 2		Set 3		Set 4	
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	8.0	6.77	8.0	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	9.28	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4.0	3.10	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









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.

ANSCOMBE’S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty 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.
Hence Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit 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.

Pearson correlation coefficient (r)	Correlation type	Correlation type	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction	When one variable changes, the other variable changes in the same direction
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

Scaling refers to the process of transforming features in a dataset to have a similar scale or range. It's a crucial preprocessing step in machine learning to ensure that all features contribute equally to the analysis and model training. If scaling is not done, then a amachine learning algorithm tends to weigh greater values higher and consider smaller values as lower values, regardless of the unit of values.

Normalized and standardized scaling have the following differences

1. Normalized scaling maps features to a specific range (usually 0 to 1), while standardized scaling transforms features to have a mean of 0 and a standard deviation of 1
2. Normalized scaling can be sensitive to outliers if the range is affected, while standardized scaling is less affected by outliers due to its reliance on the mean and standard deviation.
3. Normalized scaling retains the original distribution's shape within the specified range, while standardized scaling shifts the distribution to have a mean of 0.

the choice between normalized and standardized scaling depends on the nature of your data, the algorithm you're using, and whether you're concerned about outliers.

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.

This can happen due to one or more of the following :

- Perfect Multicollinearity : erfect multicollinearity means that one variable can be perfectly predicted from a linear combination of other variables, resulting in a rank-deficient matrix during calculations
- Low Number of Data Points: When the number of data points is very small compared to the number of predictor variables, the computation of VIF values can be unstable, potentially resulting in large or infinite values.
- Use of Dummy Variables : When creating dummy variables from categorical variables, if you include all categories without dropping a reference category, you might encounter multicollinearity issues that lead to high VIF values.

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.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.
- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis