Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on performing simple analysis of categorical variables on the dependent variable (cnt), we can come up with the initial analysis:

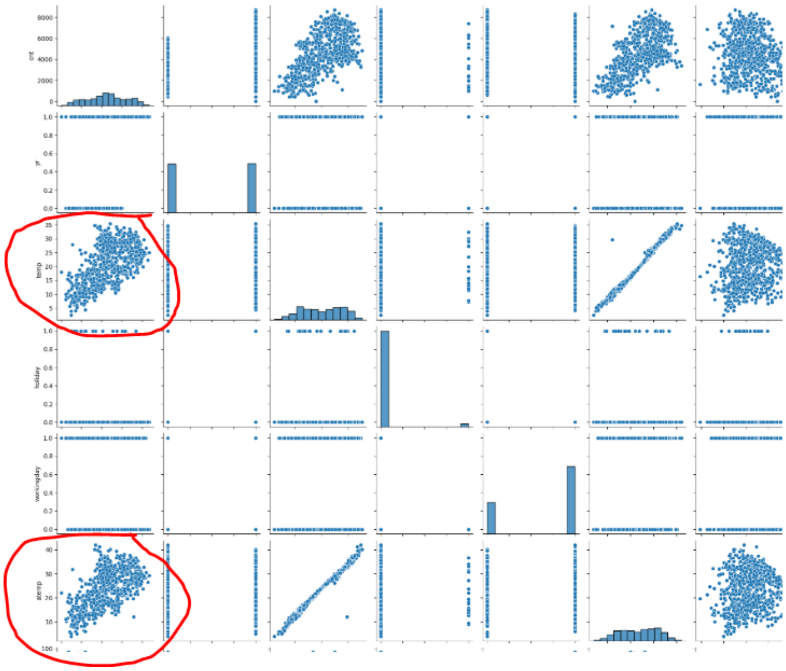
1. Season: The count of total rental bikes is highest in Fall and lowest in spring. This trend can be observed consistently across both years as well.
2. Year: The count of total rental bikes has more than doubled in 2019 compared to 2018.
3. Month: The count of total rental bikes in 2018 has been highest in the month of June while in 2019 it has been highest in the month of September.
4. Holiday: The count of total rental bikes has been highest on non-holidays
5. Weekday and Workingday: Most rental bikes were taken on a non-working Wednesday and a working Thursday. It is also evident that on weekends, we have good number of bike rentals.
6. Weathersit: It is clearly evident that no rental bikes have been taken when the weather was heavy rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. For the other three categories, most number of rental bikes were taken when the weather was clear.

After performing linear regression, our model includes July (month), Light Snow (weathersit) and Spring (season).

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

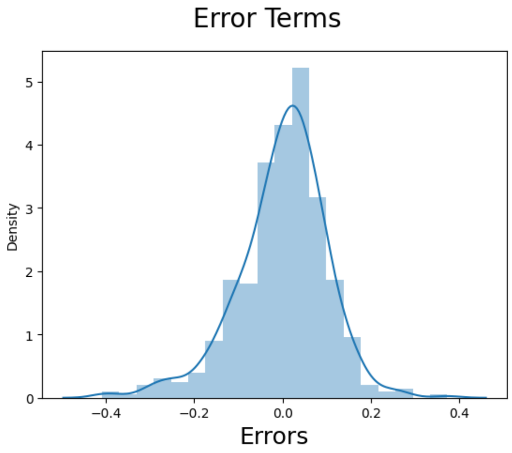
Usually when we create dummy variables from a categorical variable, the number of columns that would be created is equal to the number of unique values in the categorical variable. Since we will be able to infer one of the values from the remaining values, we would not require one column and thus we use drop_first=True to remove the first column. This in a way is done to reduce the multicollinearity between the columns.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



Temp and atemp have the highest correlation with the target variable cnt

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)



The distplot of error terms when plotted is a normal distribution curve with mean centered around 0. These are the assumptions that I have checked against to test linear regression.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top three features per my model are the following (considering the coefficients):

1. Temp: With a coefficient of ~0.5, a unit increase in temp will increase the demand in bike rentals by 0.5 (considering other variables as constant).
2. Year: With a coefficient of ~0.24, a unit increase in year will increase the demand in bike rentals by 0.24 (considering other variables as constant).
3. Hum: With a coefficient of ~-0.22, a unit increase in humidity will decrease the demand in bike rentals by 0.22 (considering other variables as constant).

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is the most basic supervised machine learning model. This is primarily used in predicting variables that are continuous in nature for example, predicting 12th marks based on 10th marks. It is based on the equation Y = mX+c

m is referred to as slope or tan of angle the line makes with the y-axis and c is y-intercept, the value of y when x is zero.

This line describes the best fit between the variables X (predictor variable) and Y (dependent variable) with least error.

Linear regression is of two types simple (one predictor variable) and multiple (more than one predictor variable). Multiple Linear Regression follows the equation Yi = B0 + B1X1 + B2X2 + … + BiXi

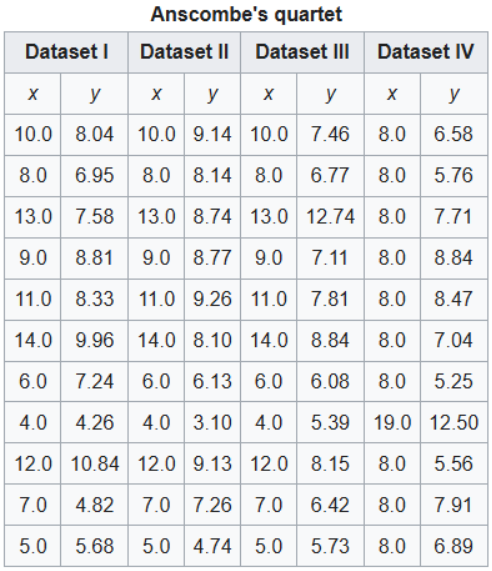
Where Bi is the coefficient of Xi and B0 is the y-intercept.

We follow the below steps in linear regression:

1. Read, understand and visualize the data
2. Prepare the data by train-test split, rescaling, dummy variables, etc.
3. Train the model
4. Perform Residual Analysis
5. Predict and Evaluate the model

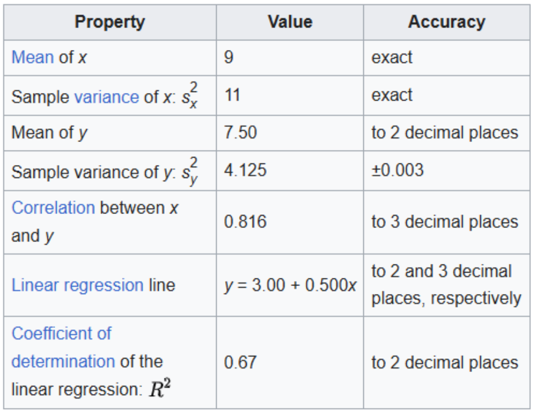
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet comprises 4 datasets (of two variables) shown below that have nearly identical descriptive statistics data (mean, variance, correlation, linear regression line, etc.).



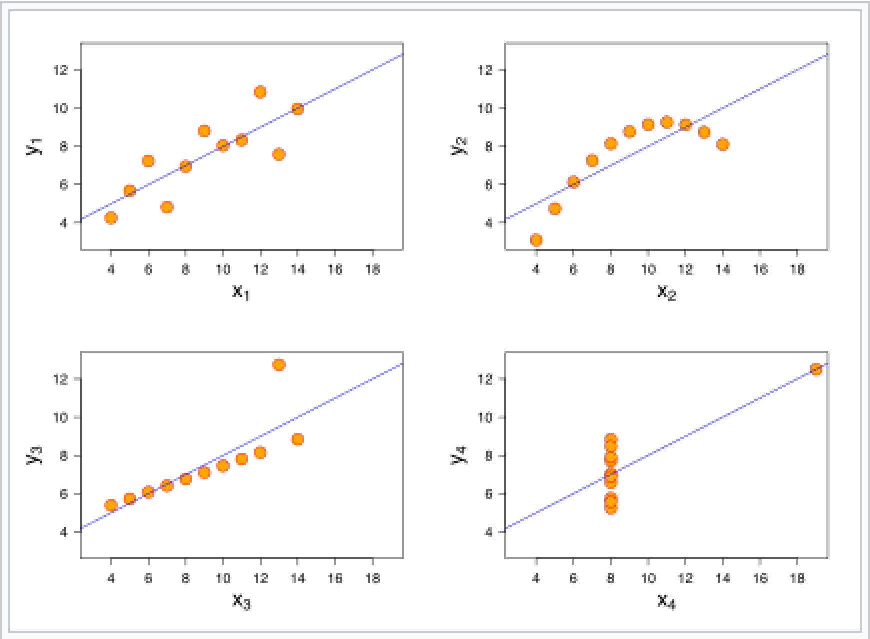
Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

While the statistics data is nearly identical, the data when visualized appear to be different. This dataset was created to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R or Pearson's Correlation Coefficient is a measure of linear correlation between two sets of data. Its value always lies between -1 and 1. Basically, Pearson's coefficient measures if a straight line can pass through the points (x,y) when plotted on a graph.

If r = 1, implies perfect linear correlation with positive slope i.e. all points can be plotted on a straight line with value of X increasing with increasing value of Y

If r = 0, implies no linear correlation

If r = -1, implies perfect linear correlation with negative slope i.e. all points can be plotted on a straight line with value of X decreasing with increasing value of Y (or vice versa)

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

When performing linear regression, we may encounter variables that are in range of 0-5 and variables that are in range of 1000-10000. In such cases, their coefficients would be disproportionate in comparison. For better interpretation and optimization, we need to bring the variables at a comparable scale. This is referred to as scaling.

We have two methods of scaling – normalized and standardized.

Normalized scaling is preferred when our data doesn’t have outliers and follows the min-max scaling approach. Once scaled via this approach the data range usually is between 0 and 1.

Standardized scaling is done when our data is distributed Gaussianly. There is no preset data range in this approach.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF is calculated as 1/(1-R**2).

When there is perfect correlation, R=1 and thus VIF will be infinity (one independent variable being explained by the other independent variables). This could also happen when we are considering a lot of variables which increases chances of multicollinearity.

So whenever we encounter such high VIF values, we need to eliminate variables that add little or no value to the linear regression model. This would result in VIF values reducing.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot short for Quantile-quantile plot is a scatter plot that compares the quantiles of two distributions. In linear regression, we could use q-q plot to check if the residuals are normally distributed and have a constant variance, key assumptions that needs to be checked. We create a Q-Q plot for residuals and compare them with the normal distribution.

Q-Q plot could have limitations when the sample size is small and in also how they are interpreted as that can be subjective and thus histograms are preferred.
