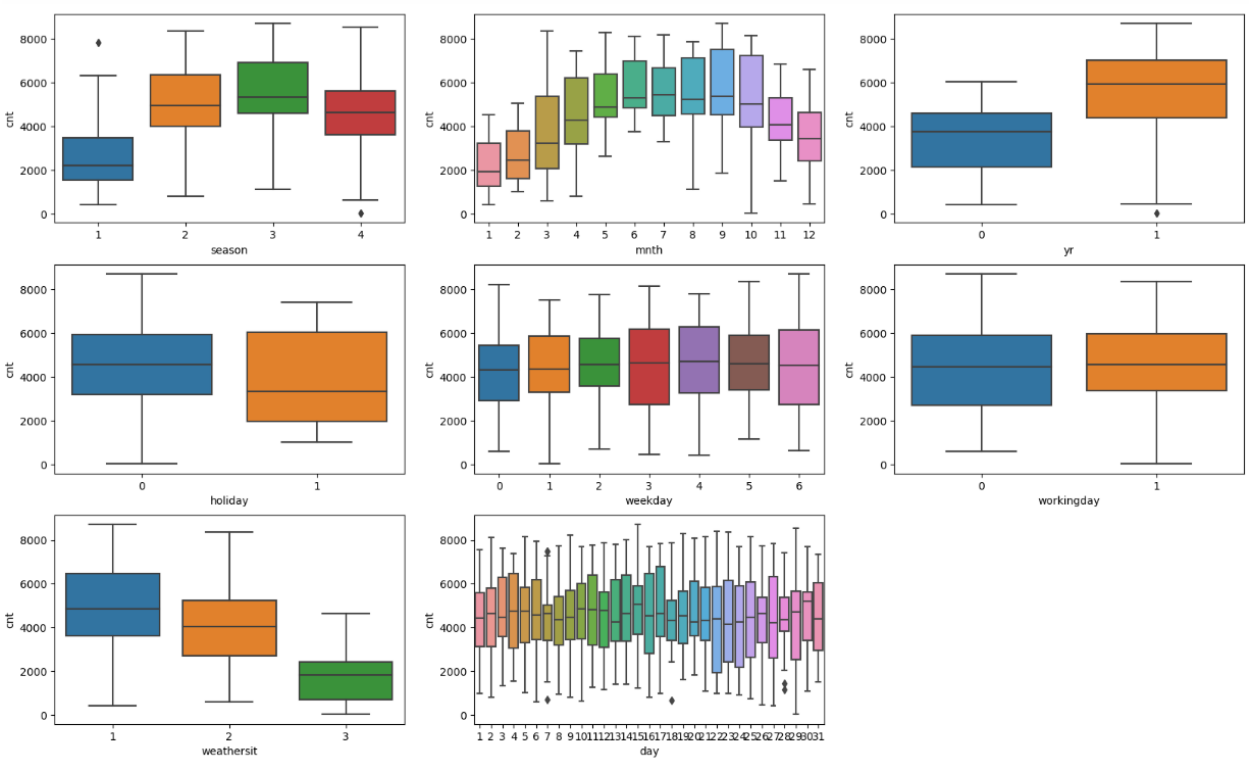
Larry Chuon

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what would you infer about the effect on the dependent variables?



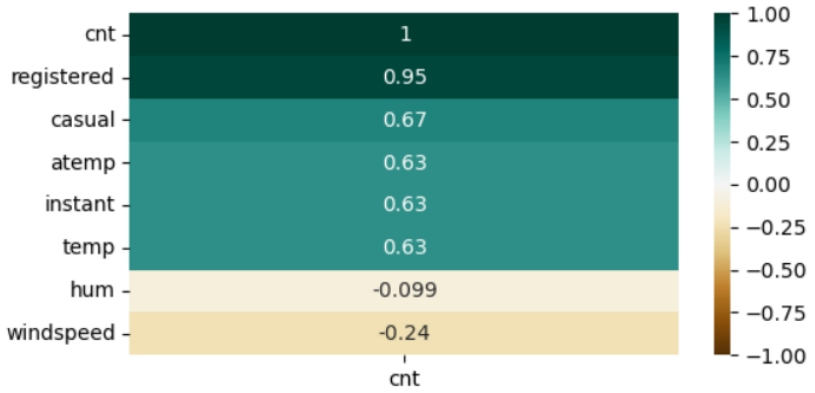
   a. Falls has the highest rental. Summer is the next highest. Both seasons are when kids are out of school. People spend time with them families and friends.
   b. Month: More bike rentals occur between May and September
   c. Year: 2019 bike rental is higher than 2018 suggesting the trend it up.
   d. Holiday: non-holiday (0) has a wider range compared to holiday
   e. Weekday: Wednesday and Thursday appear to have more bike rental
   f. Workingday: inconclusive – they are about the same
   g. Weathersit: Bike rental is high when it is Clear, Few clouds, Partly cloudy, Partly cloudy
   h. Day: It seems that bike rental occurs more in the middle of the month

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

```
DUMMIES = ['weathersit']
DUMMY_COLS = pd.get_dummies(bike_df[DUMMIES], drop_first=True)
bike_df = pd.concat([bike_df, DUMMY_COLS], axis=1)
bike_df = bike_df.drop(DUMMIES, axis=1)
bike_df.head()
```

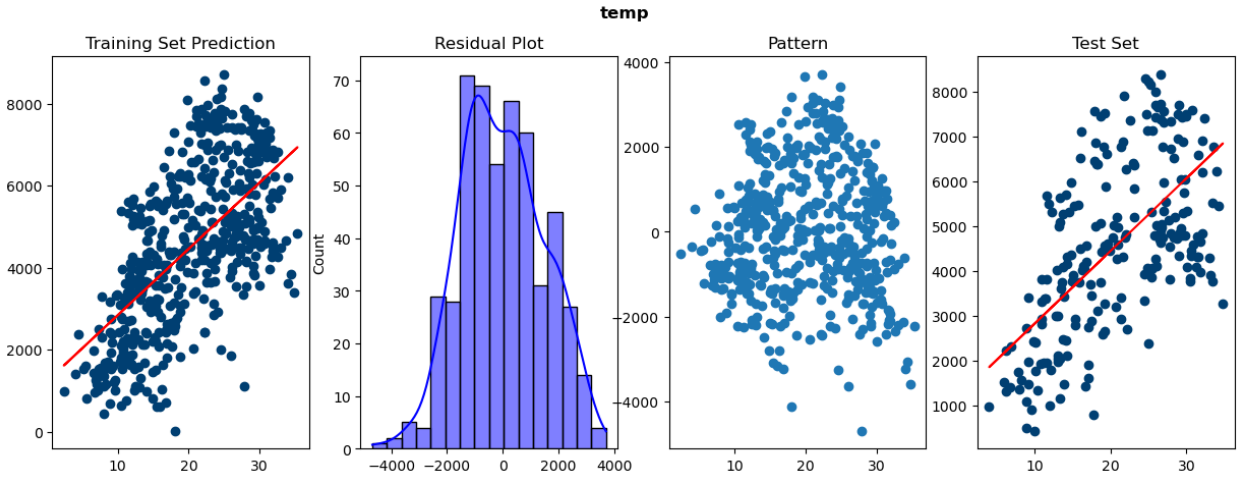
drop_first is important to use when using dummy variables, as it helps in reducing the extra column created during dummy variable creation. The remaining columns become linearly independent. As a result, it makes it easier to interpret the coefficients of the fitted model.

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



Our target variable is 'cnt'. Based on the correlation above, registered has the highest correlation with 'cnt'. Heatmap confirms the same, but I prefer looking at it this way as it is a lot clearer.

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



a. As the above showed, I plot the residual to see distribution. Then, I plot the error to determine whether is any recognizable patterns. Then, I plot the test set.
b. After each training, I evaluate the VIF, P-values, and R-squared. If the VIF score is above 5, it indicates multicollinearity. If the P-value is above 0.005, it is highly insignificant so I reiterate and drop another variable.
c. Next, I check out the r2_score.

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.806
Method:                 Least Squares   F-statistic:                     235.6
Date:                Sat, 10 Dec 2022   Prob (F-statistic):          1.45e-173
Time:                        16:51:15   Log-Likelihood:                 461.33
No. Observations:                 510   AIC:                            -902.7
Df Residuals:                     500   BIC:                            -860.3
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1957      0.031      6.395      0.000       0.136       0.256
season         0.1361      0.013     10.497      0.000       0.111       0.162
yr             0.2332      0.009     26.242      0.000       0.216       0.251
holiday       -0.0858      0.028     -3.060      0.002      -0.141      -0.031
weekday        0.0438      0.013      3.319      0.001       0.018       0.070
temp           0.4793      0.021     22.413      0.000       0.437       0.521
hum           -0.1068      0.041     -2.630      0.009      -0.187      -0.027
windspeed     -0.1581      0.028     -5.650      0.000      -0.213      -0.103
weathersit_2  -0.0550      0.011     -4.797      0.000      -0.077      -0.032
weathersit_3  -0.2522      0.029     -8.736      0.000      -0.309      -0.195
==============================================================================
Omnibus:                       57.319   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.805
Skew:                          -0.652   Prob(JB):                     1.18e-25
Kurtosis:                       4.924   Cond. No.                         19.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
""",
       Features    VIF
0        season   3.43
1            yr   2.01
2       holiday   1.04
3       weekday   3.00
4          temp   7.82
5           hum  13.61
6     windspeed   3.63
7  weathersit_2   2.08
8  weathersit_3   1.16)
```

As the chart above showed, ‘season, ‘yr, and ‘temp’ contribute significantly toward the demand of the shared bikes. There are other factors such as trend and health can potentially contribute as well. 2019 showed an increase in Bike rental, but two years’ worth of data might not suffice. More studies should be conducted.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression can explore how one or more predictor variables can be associated with one or more target variables. It can be useful in a lot of contexts. If we know the X and Y coordinates of two points, we can draw a line between them and determine the coordinates of any other point on the line. This function is called a linear function. Linear regression is the simplest form of regression analysis and can predict a single value (such as the output of a machine) based on one or more predictor values. Multiple linear regression is called multivariate linear regression if there are over one input variables. They describe the relationship between the two variables in the line. The percentage of money spent at the mall is determined by the number of customers who frequent the mall. The value of the dependent variable, y, is also increasing as the value of the independent variable, x, increases. A good line is the optimal fit for straight lines. Based on the data, we plot a line that works best for them. Linear Regression is $y = mx + b$

y = dependent variable
x = independent variable
b = slope of the line

An example of a use case for linear regression is predicting bike rental, house price or test score based on amount of time spent studying.

2. Explain the Anscombe's quartet in detail.

The Anscombe's Quartet is a set of four datasets nearly identical in simple descriptive statistics, but which contain some strange features which fool the regression model if built. If you have a specific distribution, a scatterplot is a good way to see if you have what it takes to qualify for Mensa. It's important to visually look at the data before applying an algorithm to it. Outliers are important. They represent information that is not typical or representative of the data set. To identify and handle these outliers, plot the data using techniques such as box plots and density plots.

If you're going to use linear regression, you need to understand that it's used to find the relationship between a dependent variable and an independent variable. It should analyze only such data, not to solve any other problems. A statistical model which plots all these four datasets together is called a "tent" or "parabola". A statistical model that plots each dataset separately is called a "scatterplot".

There are four datasets:

Dataset 1: This fits the linear regression model pretty well.

Dataset 2: This is a perfect fit for the quadratic regression model. Dataset 3 & 4: contains the outliers in the dataset, so linear regression is not appropriate here.

There are many reasons the data is nonlinear. Among them is that it's a time series data. The number of reviews for a product in a category differs from one product to another. Machine learning is based on algorithms and statistics, and the best ones can find patterns in your data.

When you're dealing with outliers, it pays to know what things to expect and how to manage those situations. Filtering or removing outliers from your data is an example of how a regression algorithm can be fooled. Any kind of regression algorithm could be fooled by bad data visualization. Thus, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them, which will help to make a good fit model.

3. What is Peason's R?

The Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

The correlation coefficient was planned by a 19th century English scientist named Francis Galton.

The correlation coefficient is a measure of the strength of the relationship between the two variables.

The correlation coefficient is a number that describes the strength of the relationship between the two variables.

The model doesn't consider the non-linear relationship between the two variables. It also cannot differentiate between dependent and independent variables.

How is the Correlation coefficient calculated?

The correlation coefficient helps us to understand the relationship between the two variables. If you've got an idea for a new product, you can calculate your potential ROI. For example, the growth rate of height in children increases as they get older. If we want to know the relationship between two variables, we need to collect data from a sample population.

Let's calculate the correlation between two variables, X and Y, using the Pearson's Correlation Coefficient (r).

a) 1.There are three criteria to use to assess the correlation coefficient:
b) 2. Scale of measurement should be an interval or a ratio
c) Variables should be approximately normally distributed

d) The association should be linear

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

   Scaling is changing the independent variable to make it easier to interpret by computer. It's important for your algorithm to work fast and efficiently, which will help speed up the calculations.

   Data analysis is often performed using machine learning techniques. These techniques are used to find an underlying model that fits or explains the collected data and thus reduce its dimensionality. The unit of measurement of the inputs to the model is important, and not scaling. If you don't scale your data, your algorithm will not perform well. If we want to fix this problem, we need to scale all the variables to the same level of magnitude. What is the difference between scaling and standardization? Scaling just changes the values of the parameters that the model fits, but standardizing just changes the values of the variables. There are different of scaling:

   - Normalization/Min-Max Scaling
   - Standardization Scaling

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

   If there is a simple correlation, then VIF = infinity. This means that there is a simple correlation between the two independent variables. If the correlation is perfect, the number is infinite. We need to drop one predictor to fix this problem. A large VIF suggests that they strongly tied in the corresponding variable with other variables, and thus should not be included in the regression analysis. If the VIF is high, this means that the model coefficients are very correlated, which can lead to multicollinearity problems. This means that the standard error of the coefficient is twice as high as it should be. A high value for VIF indicates the model is suffering from multicollinearity, which might cause the coefficient to be insignificant.

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

   You can use a QQ plot to compare the shapes of distributions. A Q-Q plot is a simple way of analyzing your data. The first percentile is the actual distribution of your data, the second one is the variable (distribution) that you are testing your hypothesis for. It is not likely that the difference between the quantiles will be exact. The closer the values are, the more similar they are. Quantile regression is a method used to analyze data that is useful for understanding how the relationship between two variables changes across different quantiles of the variable in the higher range. Quantiles are fractions of the population. The median is a better way to calculate an average than the mean.