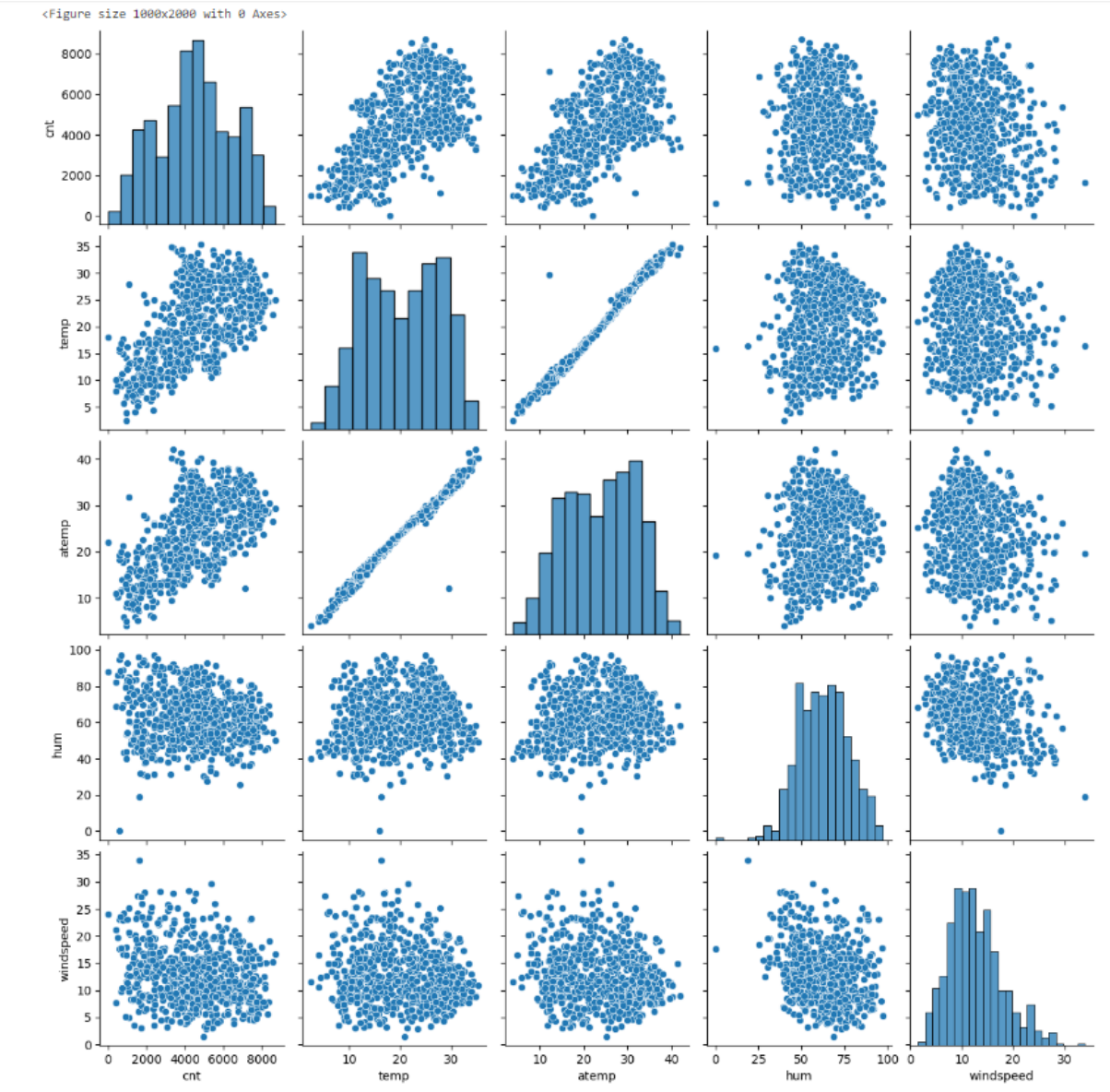
Assignment-Based Subjective Questions

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Atemp, temp had high correlation between them and had linear relationship with cnt variable



Note: Could not complete the full analysis as there were errors generating heatmap. Tried all sources on Google, but could not find the “Key Value Error” with corr().

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

As per the thumb rule, we need to take n-1 levels each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels.

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

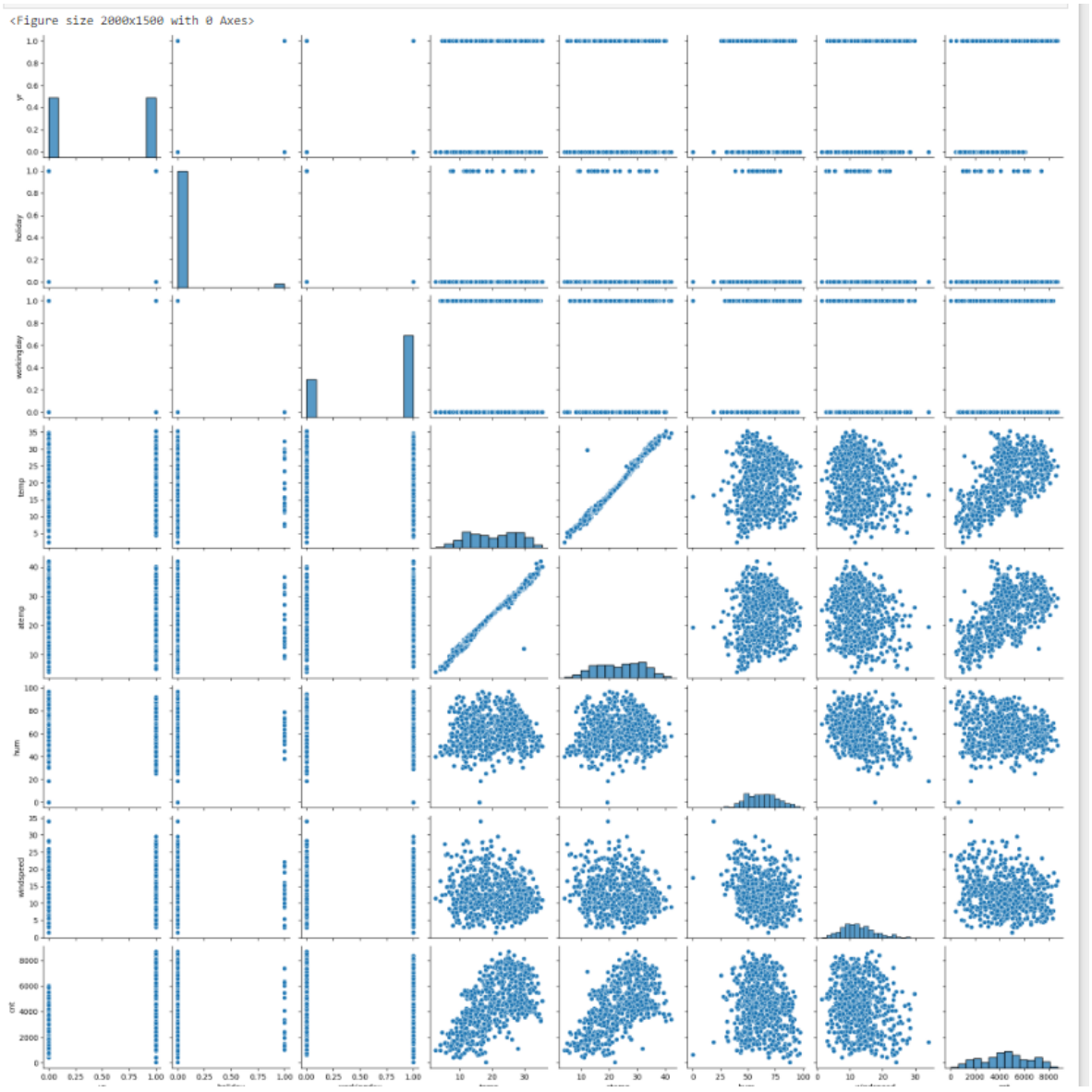
'atemp' and 'temp' variables have highest correlation with 'cnt'

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

Linearity,,Homoscedasticity, and Multicollinearity are used to validate the assumptions

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

Note: I could not build the full model due to Python error. Looking at the plots, I could infer – temperature, year and season



General Subjective Questions

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

Linear regression is a form of predictive modeling technique which explains the relationship between the dependent (target variable) and independent variables (predictors). The linear relationship is expressed in the form of y=mx+c where m is gradient and c is the intercept. With this kind of equation and relation, importance of dependent varaiables can be determined. If there are multiple variables, their correlation with the target variable is used to determine the best possible independent variable among others. The best fit line is used to identify the closest variable. O

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that tricks the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.There are these four data set plots which have nearly same statistical observations, which

provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

3. What is Pearson's R? (3 marks)
   Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

Scaling is used to transform data so that it fits within a specific scale. It one of the data pre-processing steps where data is set in specific scale and bring uniformity in the analysis.. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.

2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.

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

VIF(VarianceInflationFactor) basically explains the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables.

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

   Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

   The linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

   Importance of QQ Plot in Linear Regression :

   In Linear Regression with a train and test dataset we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check