0.0.1 Question 1c

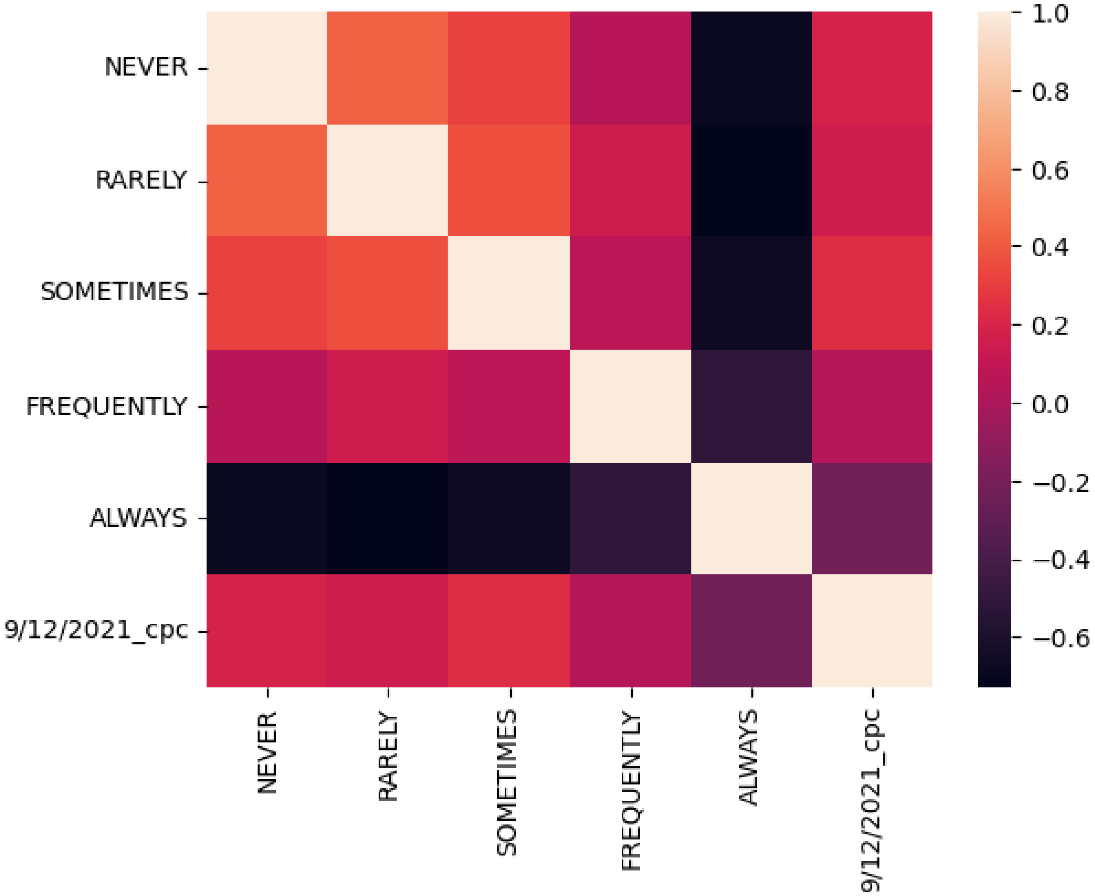
Our goal is to use county-wise mask usage data to predict the number of COVID-19 cases per capita on September 12th, 2021 (i.e., the column `9/12/2021_cpc`). But before modeling, let's do some EDA to explore the multicollinearality in these features, and then we will revisit this question in part 4.

Create a visualization that shows the pairwise correlation between each combination of columns in `mask_data`. For 2-D visualizations, consider Seaborn's heatmap. Remember to add a title to your plot.

Hint: You should be plotting 36 values corresponding to the pairwise correlations of the six columns in `mask_data`.

```
In [9]: sns.heatmap(mask_data.corr())
```

```
Out[9]: <AxesSubplot:>
```


NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS
9/12/2021_cpc
NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS
9/12/2021_cpc
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

0.0.2 Question 1d

(1) Describe the trends and takeaways visible in the visualization of pairwise correlations you plotted in Question 1c. Specifically, how does the correlation between pairs of features (i.e. mask usage) look like? How does the correlation between mask usage and cases per capita look like?

(2) If we are going to build a linear regression model (with an intercept term) using all five mask usage columns as features, then what problem will we encounter?

(1) It's clear that with the increased frequency of mask wearing, the correlation with covid chance drops to lower and lower. Also, 'NEVER', 'RARELY' and 'SOMETIMES' correlate well, but not with 'FREQUENTLY', and very poorly correlate with 'ALWAYS'.

(2) I think these categorical features are not very quantitative. They will lead to underfitting and poor prediction results.

0.0.3 Question 2b

To visualize the model performance from part (a), let's make the following two visualizations:

(1) The predicted values vs. observed values on the test set,

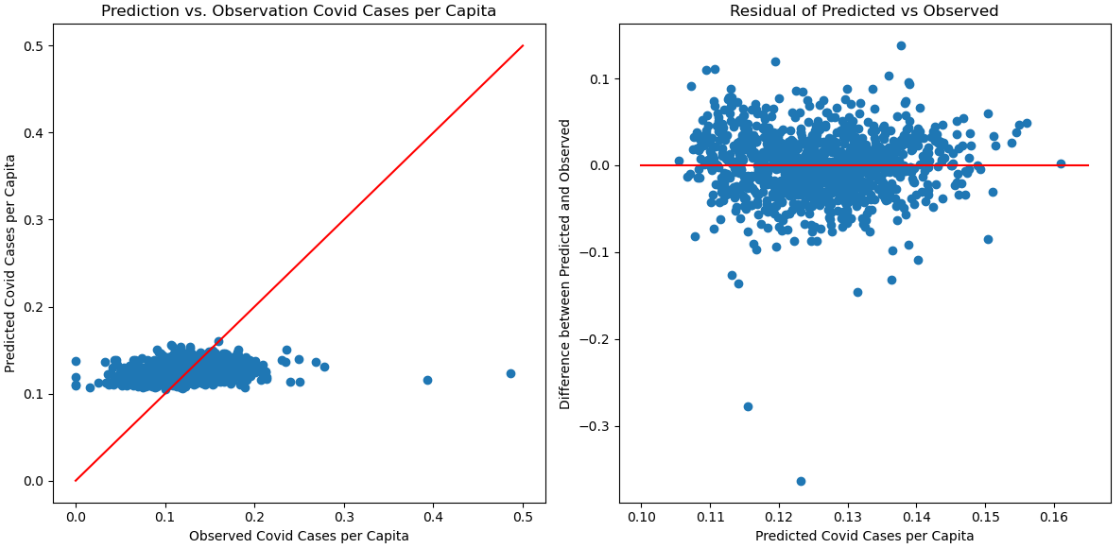
(2) The residuals plot. (Note: in multiple linear regression, the residual plot has predicted values vs. residuals)

Note: * We've used `plt.subplot` (documentation) so that you can view both visualizations side-by-side. For example, `plt.subplot(121)` sets the plottable area to the first column of a 1x2 plot grid; you can then call Matplotlib and Seaborn functions to plot that area, before the next `plt.subplot(122)` area is set. * Remember to add a guiding line to both plot where $\hat{y} = y$, i.e., where the residual is 0. * Please add descriptive titles and axis labels for your plots!

```
In [12]: plt.figure(figsize=(12,6))      # do not change this line
         plt.subplot(121)                # do not change this line
         # (1) predictions vs. observations
         plt.scatter(y_test, y_pred_test)
         plt.plot([0,0.5],[0,0.5],'r')
         plt.xlabel("Observed Covid Cases per Capita")
         plt.ylabel("Predicted Covid Cases per Capita")
         plt.title("Prediction vs. Observation Covid Cases per Capita")

         plt.subplot(122)               # do not change this line
         # (2) residual plot
         plt.scatter(y_pred_test, y_pred_test-y_test)
         plt.plot([0.1,0.165],[0,0], 'r')
         plt.xlabel("Predicted Covid Cases per Capita")
         plt.ylabel("Difference between Predicted and Observed")
         plt.title("Residual of Predicted vs Observed")


         plt.tight_layout()             # do not change this line
```


Prediction vs. Observation Covid Cases per Capita
Predicted Covid Cases per Capita
Observed Covid Cases per Capita
0.0
0.1
0.2
0.3
0.4
0.5
Residual of Predicted vs Observed
Difference between Predicted and Observed
Predicted Covid Cases per Capita
0.1
0.0
−0.1
−0.2
−0.3
0.10
0.11
0.12
0.13
0.14
0.15
0.16

0.0.4 Question 2c

Describe what the plots in part (b) indicate about this linear model.

Plot b indicates that regardless of the observed covid cases, the predicted cases don't really change. It's basically a bad prediction.

0.0.5 Question 3d

Interpret the confidence intervals above for each of the θ_i, where θ_0 is the intercept term and the remaining θ_i's are parameters corresponding to mask usage features. What does this indicate about our data and our model?

Describe a reason why this could be happening.

Hint: Take a look at the design matrix, heatmap, and response from Question 1!

All the theta intervals include 0, and have a very large range. This means the estimates are not precise and could be statistically insignificant. The reason is likely because of the vague categorical classification of the mask wearing data.

0.0.6 Question 4b

Comment on the ratio `ratio`, which is the proportion of the expected square error on the data point captured by the model variance. Is the model variance the dominant term in the bias-variance decomposition? If not, what term(s) dominate the bias-variance decomposition?

Note: The Bias-Variance decomposition from lecture:

$$\text{model risk} = \sigma^2 + (\text{model bias})^2 + \text{model variance}$$

where σ^2 is the observation variance, or “irreducible error”.

From the result, the model risk is much larger compared to the model variance. So model risk dominates.

0.0.7 Question 4d

Propose a solution to reducing the mean square error using the insights gained from the bias-variance decomposition above.

Assume that the standard bias-variance decomposition used in lecture can be applied here.

Type your answer here, replacing this text.