0.0.1 Question 2c

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`.

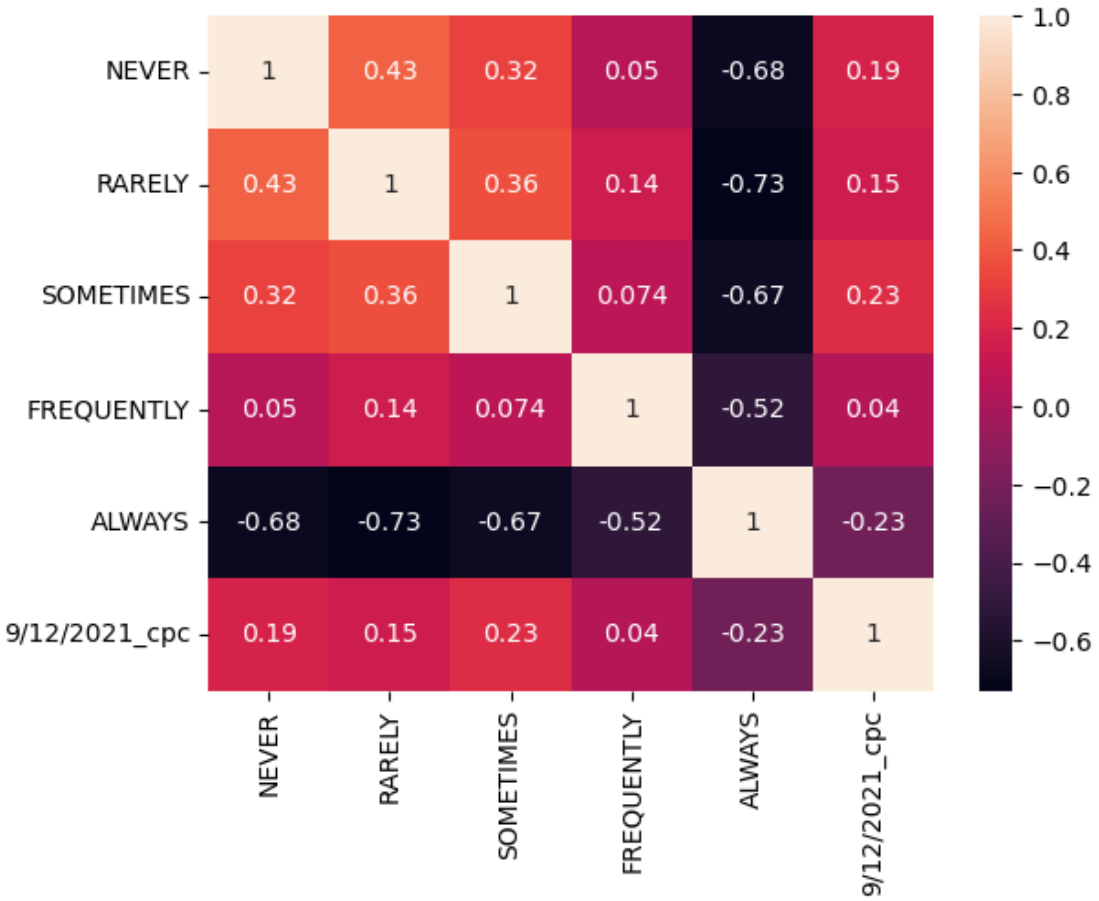
```
In [9]: mask_data.drop('NEVER', axis=1)
```

Out[9]:

	RARELY	SOMETIMES	FREQUENTLY	ALWAYS	9/12/2021_cpc
0	0.074	0.134	0.295	0.444	0.165411
1	0.059	0.098	0.323	0.436	0.152429
2	0.121	0.120	0.201	0.491	0.134003
3	0.034	0.096	0.278	0.572	0.171440
4	0.114	0.180	0.194	0.459	0.158538
…	…	…	…	…	…
3136	0.295	0.230	0.146	0.268	0.143205
3137	0.157	0.160	0.247	0.340	0.196238
3138	0.278	0.154	0.207	0.264	0.158496
3139	0.155	0.069	0.285	0.287	0.144330
3140	0.129	0.148	0.207	0.374	0.122942

[3141 rows x 5 columns]

```
In [10]: sns.heatmap(mask_data.corr(), annot=True);
```


NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS
9/12/2021_cpc
1
0.43
0.32
0.05
-0.68
0.19
0.43
1
0.36
0.14
-0.73
0.15
0.32
0.36
1
0.074
-0.67
0.23
0.05
0.14
0.074
1
-0.52
0.04
-0.68
-0.73
-0.67
-0.52
1
-0.23
0.19
0.15
0.23
0.04
-0.23
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

0.0.2 Question 2d

(1) Describe the trends and takeaways visible in the visualization of pairwise correlations you plotted in Question 2c. 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, what could be the problem?

ALWAYS feature has negative correlation with all other features includeing the response feature (9/12/2021_cpc). This could be due to the fact that people who took precaustions (wore mask always) were safer and got less affected by the virus compared to those who didn't take the virus much serious. Some mask usage feartures are strongly correlted (e.g., NEVER, RARELY and SOMETIMES) with each other, so if we chose to build a regression model the model's X_transpose X would not invert and therefore this could have become a problem if we build a linear regression model.

0.0.3 Question 3b

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)

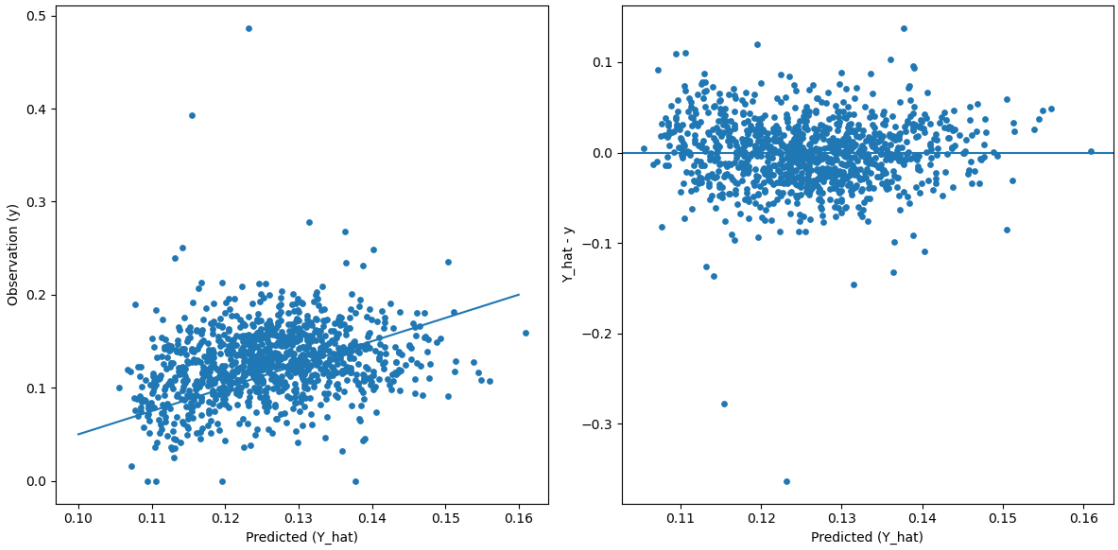
Some notes: * We've used `plt.subplot` (documentation) so that you can view both visualizations side-by-side. For example, `plt.subplot(121)` sets the plottable area to the first column of a 1x2 plot grid; you can then call Matplotlib and Seaborn functions to plot that area, before the next `plt.subplot(122)`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

```
In [13]: plt.figure(figsize=(12,6))      # do not change this line

         plt.subplot(121)                # do not change this line
         # (1) predictions vs observations
         plt.scatter(x=y_test_predicted, y=y_test, s=15, alpha=1)
         plt.xlabel('Predicted (Y_hat)')
         plt.ylabel('Observation (y)')
         plt.plot([.1,.16],[.05 ,0.2])

         plt.subplot(122)               # do not change this line
         # (2) residual plot
         plt.scatter(x=y_test_predicted, y=y_test_predicted-y_test, s=15, alpha=1)
         plt.xlabel('Predicted (Y_hat)')
         plt.ylabel('Y_hat - y')
         plt.axhline(0)

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


Observation (y)
Predicted (Y_hat)
Y_hat - y
Predicted (Y_hat)

0.0.4 Question 3c

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

Although the RMSE is really small, but there is a clear pattern in the residual plot. In other words, the points are not evenly distrubuted aroud the line (y=0) in the residual plot. The middle points are mostly under the line while the in the two ends they are above the line. This means we may have to transform some features to build a better model instead.

0.0.5 Question 4d

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 2!

We know that our design matrix and the response (cases per capita) vector cotains small values. So, this means the weights for the design matrix also must be small or close to 0. I believe this would be one reason that theta_i's contain 0. Another reason would be that some mask usage features may not be useful to our model to predict case per capita. This is why the weights for these features include 0. This also make sense because if we look at heatmap we can clearly see that almost all mask usage features have weak correlation with case per capita.

0.0.6 Question 5b

Comment on the ratio `prop_var`, 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".

The model variance is not the dominant term in the bias-variance decomposition because the prop_var value is very small. On the other hand observation variance (segma^2) or model bias are the two terms that dominate to the bias-variance decomposition.

0.0.7 Question 5d

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.

As stated in part 5b, model bias contributes more to the mean sqaure error as the value of prop_ratio was very small. So, in order to reduce the sqaure error we have to increase the model complexity to reduce the bias. By adding useful features we capture as close the true relationship as possible, however, we never capture the true relationship due to the noise in the model which we can't do anything about it.