0.0.1 Question 1c

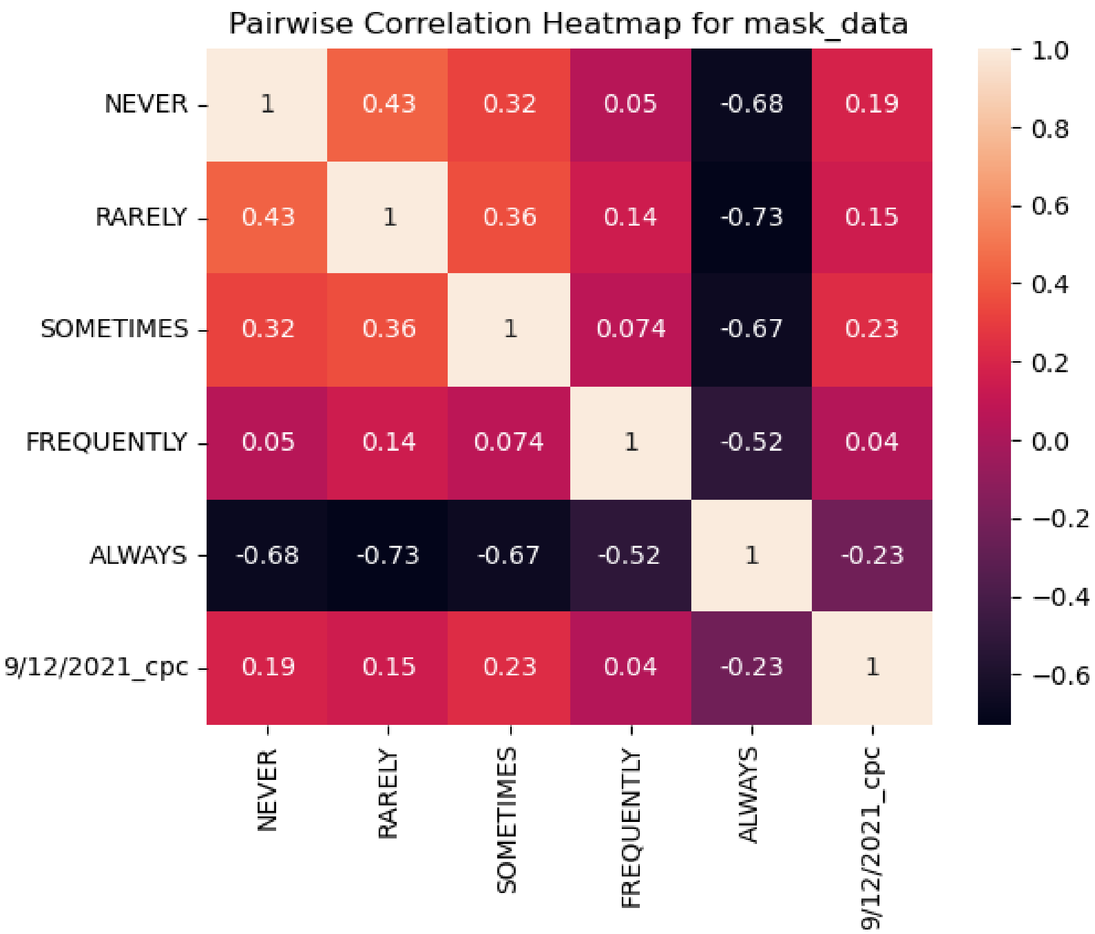
Our goal is to use county-wise mask usage data to predict the number of COVID-19 cases per capita on September 12th, 2021 (i.e., the column `9/12/2021_cpc`). But before modeling, let's do some EDA to explore the multicollinearity in these features, and then we will revisit this in question.

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 title your plot.

Hint: You should be plotting 36 values corresponding to the pairwise correlations of the six columns in `mask_data`. You may optionally set `annot=True`, but it isn't necessary.

```
In [53]: sns.heatmap(mask_data.corr(), annot=True)
         plt.title("Pairwise Correlation Heatmap for mask_data")
```

```
Out[53]: Text(0.5, 1.0, 'Pairwise Correlation Heatmap for mask_data')
```


Pairwise Correlation Heatmap for mask_data
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 1d

Describe the trends and takeaways visible in the visualization of pairwise correlations you plotted in Question 1c. Specifically, what does the correlation between pairs of features (i.e., mask usage categories) look like? What does the correlation between mask usage categories and COVID-19 cases per capita look like?

Correlation becomes stronger from always to never. And we see a negative correlation between cases per capita and always column.

0.0.3 Question 1e

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

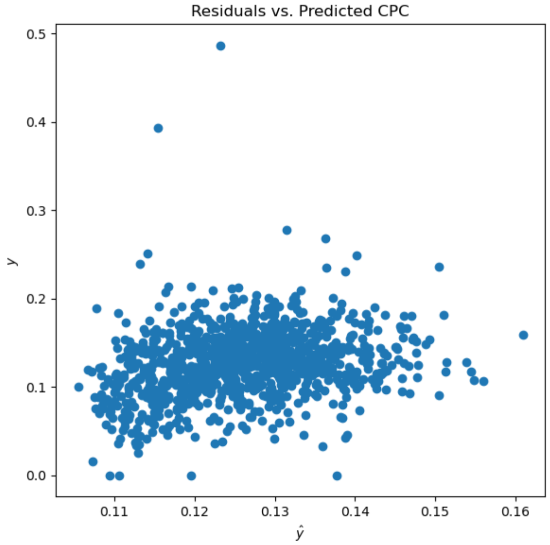
Multicollinearity, because our independent variables are highly correlated with each other. The five mask usage highly correlated because they add up to a constant sum.

0.0.4 Question 2b

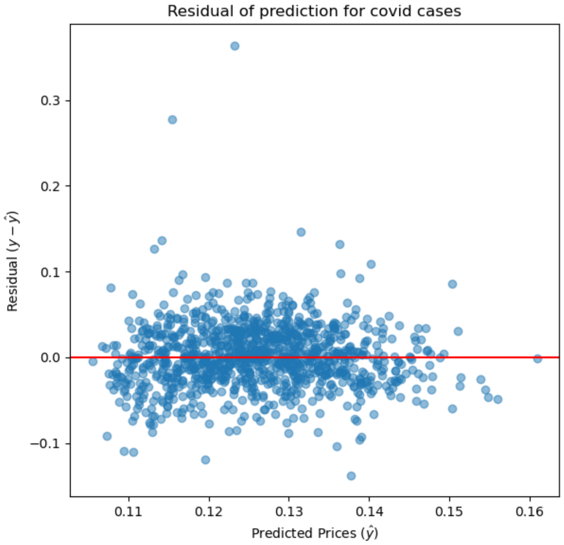
To visualize the model performance from part (a), let's make the following two visualizations: 1. The predicted values vs. observed values on the test set, 2. The residuals plot. (Note: in multiple linear regression, the residual plot has the residuals plotted against the predicted values).

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 plots where $\hat{Y} = Y$, i.e., where the residual is 0.** * Please add descriptive titles and axis labels for your plots!

```
In [56]: plt.figure(figsize=(12,6))      # do not change this line
         plt.subplot(121)                # do not change this line
         # 1. plot predictions vs. observations
         plt.scatter(Y_pred_test, Y_test)
         plt.ylabel("$y$")
         plt.xlabel("$\hat{y}$")
         plt.title("Residuals vs. Predicted CPC")

         plt.subplot(122)               # do not change this line
         # 2. plot residual plot

         plt.scatter(Y_pred_test, Y_test- Y_pred_test, alpha=0.5)
         plt.ylabel("Residual $(y - \hat{y})$")
         plt.xlabel("Predicted Prices $(\hat{y})$")
         plt.title("Residual of prediction for covid cases")
         plt.axhline(y = 0, color='r');

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



Residuals vs. Predicted CPC



Residual of prediction for covid cases

0.0.5 Question 2c

Describe what the plots in part (b) indicate about this linear model. In particular, are the predictions good?

Yes, the predictions are good. The residual plot shows uniform distribution with a few outliers.

0.0.6 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!

0 is included in the 95% confidence interval for every parameter of the model. We are uncertain that any of the input variables impact the response variable. This is due to response variables are highly correlated with each other.

0.0.7 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 the lecture:

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

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

The ratio is very small. This implies that the model bias and the error are likely the dominant factors contributing to the expected square error on the given data point.

0.0.8 Question 4d

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

Assume that the standard bias-variance decomposition used in Lecture 19 can be applied here.

Use k-fold cross-validation, to estimate the model's performance on unseen data. Cross-validation helps assess how well the model generalizes to new data and can guide decisions on model complexity and regularization.