Introduction

This report creates a regression analysis on a random sample of 60 out of 578 census blockgroups in the "SF Sea Level Rise, Flooding, and Health" dataset, with an aim to create the highest predictive model of flood vulnerability to inform policymakers about the allocation of flood protection resources. Research shows that flood vulnerability is highly determined by geographical factors (Ganová, L., 2011), which informs my research question: "Is there a correlation between elevation level and FloodHealthIndex in the population of 578 census blockgroups?". The report starts with data statistics, correlation visualization to condition checking and linear regression building, then evaluation of slope statistical significance and expansion of multi-regression model.

Dataset

The dataset contains 17 independent variables and 1 dependent variable (FloodHealthIndex). To maximize prediction power, I evaluate the relationship between Elevation - the most correlated variable with the response (Pearson's r = - 0.7) (Appendix A).

	Explanation [1]	Type
Independent variable	**FloodHealthIndex** FloodHealthIndex is the result accumulation of Social and Demographic, Exposure, Health and Housing	Both variables are continuous quantitative variables. Given a finite interval, continuous variables can assume infinite

[1] #variables: I explained the variables in terms of their meaning, measurement, unit, as well as how we should interpret the quantity of the result with implication to real-life implication. For example, since elevation has a negative value, the way that I interpret it would be the minimum feet below the sea level. I justified the type of 2 variables using multiple methods: the values within an interval, the equation of those and the actual observed values.

	Vulnerability indicators of a census blockgroup. The higher the index (the unit is point), the more vulnerable an area.	values in that interval. For FloodHealthIndex, the mathematical equation deals with the average/ division of other variables. For Elevation, the feet unit result does not take a clear preceding and succeeding order. Both variables take 4-5 decimal places. Therefore, they are continuous.
Dependent variable	**Elevation** Elevation is measured by the minimum elevation in feet above sea level. The lower the elevation, the closer to sea level and the riskier an area.	

Table 1: Summary statistics for 2 variables (Appendix B)		
	x (Elevation)	y (FloodHealthIndex)
Count	$n1 = 60$	$n2 = 60$
Mean	$\bar{x}1 = 105.748785$	$\bar{y} = 52.337878$
Standard Deviation	$sx = 95.115572$	$sy = 13.713759$
Range	$[-16.99; 433.669]$	$[26.2998; 81.1858]$

Methods

To start fitting the linear regression between 2 variables, I check the LINER conditions:

Checking conditions

L - Linearity [2]

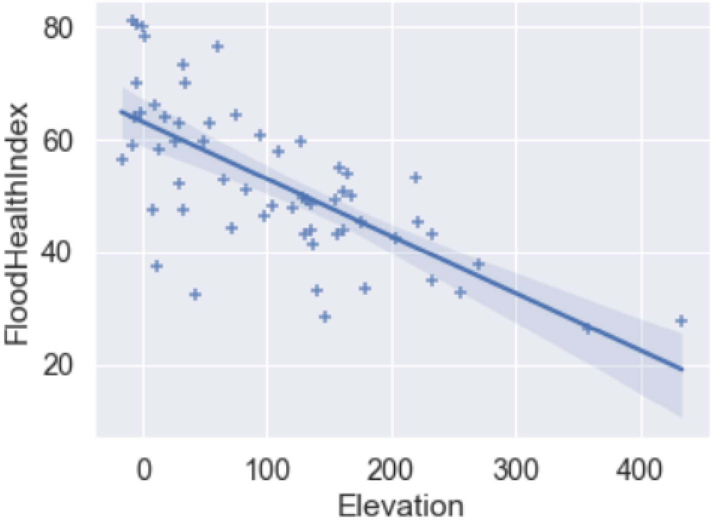


Figure 1. Scatter plot of elevation (predictor) and FloodHealthIndex (response)

The data was fitted to a straight line, indicating a strong and negative correlation between elevation and FloodHealthIndex. An increase in elevation correlated with a decrease in FloodHealthIndex.

I - Independence

The dataset was retrieved from 60 out of 578 census blockgroups. It is larger than the condition (< 10% of the population = 57 census blockgroups), in which removing each observation possibly changes the probability of the next draw in a finite population and causes an overestimation of the standard error. Therefore, I apply the Finite Population Correction Factor to the standard error of the slope: SE(b1) corrected = SE (b1) * $\sqrt{\frac{\boldsymbol{N-n}}{\boldsymbol{N-1}}}$ (Appendix ..)

[2] # dataviz: For each condition, I used a specific kind of data visualization to effectively convey the right information. The linear condition is best evaluated when looking at the scatter plot, as well as to establish an initial intuition about the relationship between 2 variables. According to the Central Limit Theorum, distribution of (sufficiently large) random samples from the population would be normal. The same applies for residuals. Therefore, histogram is the best method to evaluate normal distribution of residuals by looking at clear, absurd outliers.

N - Normality residuals

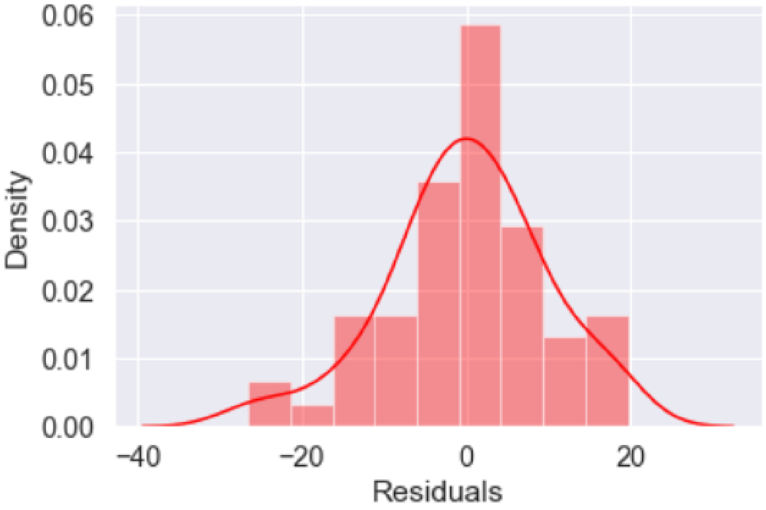


Figure 2. Histogram of residuals of Elevation and FloodHealthIndex

The histogram of residuals shows a normal distribution, most values falls around 0 even though the histogram is slightly left skewed.

E - Equal variances

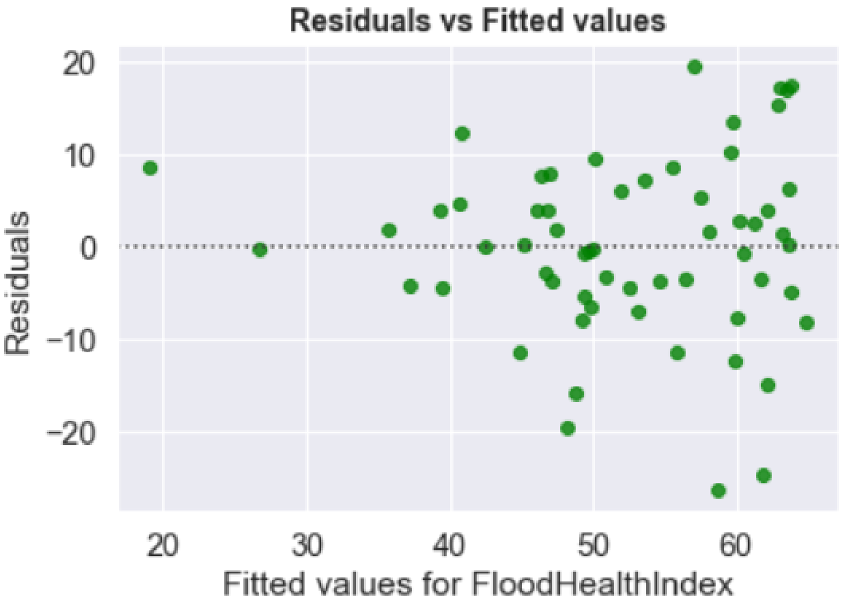


Figure 3. Residuals plot of Elevation and FloodHealthIndex

The heteroscedasticity variance of the residuals is unequal/ not constant over a range of measured values. The vertical range of residuals increases from left to right, meaning less fitted to the regression line.

R - Random

The description of the dataset indicates a random sample from the population dataset.

Overall, the scatterplot is not football-shaped because of heteroscedasticity, but it is without large outliers and shows a linear association.

Pearson's correlation coefficient, r [3]

Correlation coefficient shows how strongly the variables correlate linearly. Since it lies between ± 0.50 and ± 1 (r = - 0.703), there is a strong and negative correlation between elevation and FloodHealthIndex. When elevation increases, there is a decrease in FloodHealthIndex. However, correlation does not mean causation: elevation does not decrease FloodHealthIndex.

Regression equation

The above correlation coefficient means that the regression line fits the data more closely, not the regression line is steeper. Instead, it is expressed by the regression equation FloodHealthIndex = -0.101 * Elevation + 63.062. The slope indicates that each additional foot of the minimum elevation decreases the FloodHealthIndex by 0.101 points. The larger the slope, the steeper the regression line. The y-intercept indicates that given a census blockgroup is 0 feet above sea level, the FloodHealthIndex would be 63.062. However, this is impractical in real-life scenarios of people living on land, which leads to extrapolation since we don't have points that are close to the origin. From the coefficient of determination ($R^2 = 0.495$), we see that 49.5% of

[3] # correlation: I explained what correlation coefficient means in general and applied the concept in my report. I interpreted what the result says about the relationship between 2 variables. I am cautious that correlation does not mean causation. In conclusion, I examined the effect of the unmet conditions on the correlation coefficient, especially heteroscedasticity. I used correlation coefficient to inform my initial selection of Elevation variable, and to inform several variables selection in multiple regression model.

the variation in the FloodHealthIndex is explained by the variation in elevation. This is aligned with the correlation coefficient ($R^2 = r^2$) because this is a simple linear regression.

Significance [4]

The strength of the regression needs to be examined further for statistical significance through hypothesis testing:

Null hypothesis	H0: there is no linear relationship between FloodHealthIndex and elevation in the population; $\beta 1 = 0$
Alternative hypothesis	HA: there is some linear relationship between FloodHealthIndex and elevation in the population; $\beta 1 \neq 0$

I have no assured prediction or interest in whether there is a positive/ negative correlation between 2 variables; therefore, I use a 2-tailed test.

The consequence of committing Type II error (indicating that there exists no relationship between FloodHealthIndex and elevation while it actually does) is more serious because it misleads and exacerbates the allocation of resources in flood-prone areas. Therefore, I will set the significance level $\alpha = 0.1$ and confidence interval of 90%. I only reject the null hypothesis if it has a p-value less than 0.1, which is more probable than when $\alpha = 0.05$.

The p-value ≈ 0 ($p <$ alpha = 0.1) and the constructed confidence interval [-0.123, -0.08] (does not include 0) show that the null hypothesis is rejected and the slope coefficient is statistically significant. Firstly, this p-value represents the probability of getting a sample this extreme or more given the null is true. That means the probability of no linear relationship between elevation and FloodHealthIndex is significant and does not lie in the heavily extreme

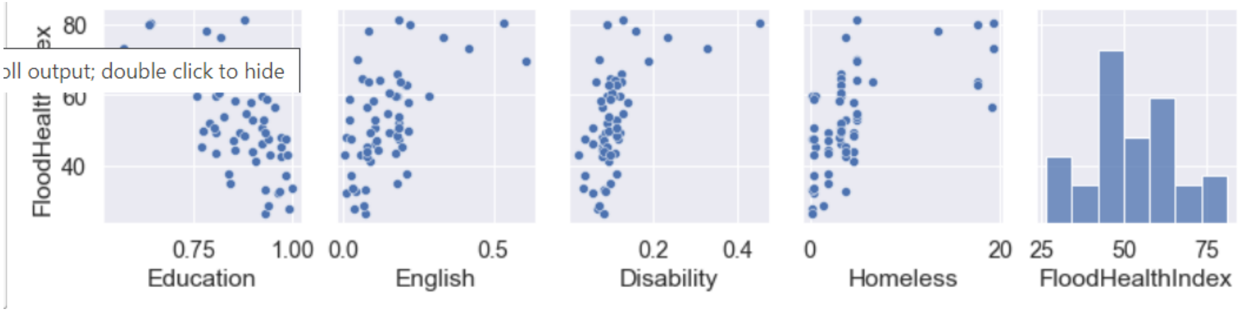
[4] # significance: I specified that the significance is of the slope coefficient, not the mean as in the previous assignment. I used both p-value and confidence interval to compliment each other and strengthen the argument. I also use p-value in forward selection to evaluate the significance of a variable when added to the model (to stop adding NonWhite). I justified why I used alpha = 0.1 and used Bofferroni correction in each selection round to account for the multiple tests.

small tails. The observed linear relationship is less likely to happen by chance or sampling variance. The result is aligned with the confidence interval that does not capture the null hypothesis. In the long run, if more samples and more slope confidence intervals are conducted, 90% of those will capture the true population slope coefficient. [5]

Forward selection

I proceed with multiple regression to increase the predictive power of FloodHealthIndex. Thus, I use adjusted R-squared instead of p-value. Since I start with Elevation, the baseline of the adjusted R-squared is 0.486.

However, I first evaluate conditions and relationships between variables to narrow down an exhausted list of 16 variables. The scatter plots (Appendix …) removed variables Education, English, Disability, and Homeless due to highly non-linear relationships with the response variable.



[5] # confidence intervals: I use the confidence interval to back up for the rejection of the null hypothesis when doing the testing with p-value, and emphasize the significance of the difference. I emphasize how the confidence intervals provide more information than the testing with p-value. I interpret what the confidence interval means using the Frequentist approach and provide a non-technical interpretation in the conclusion by assigning the lower and upper bounds to the amount of decrease in FloodHealthIndex.

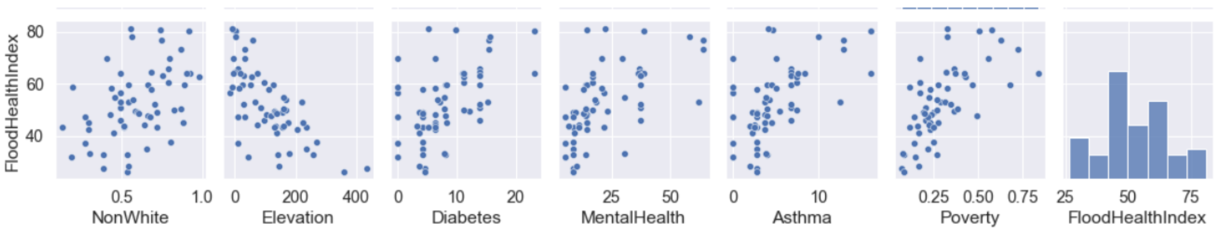


Figure 4. Scatter plots between predictors and response variables. The upper row shows non-linear relationships.

In the first round, between the remainders, the addition of Poverty leads to the highest adjusted R-squared (Rsquared_adj = 0.64) and significant p-value ($p \approx 0$) (Table 2).

Elevation (0.486)	**Poverty**	**Mental Health**	**Asthma**	**Diabetes**	**NonWhite**
Adj R squared	**0.64**	0.589	0.62	0.59	0.595

Table 2. Adjusted R-squared for 5 candidates of the first added variable

In the second round, the addition of MentalHealth leads to the highest adjusted R-squared (Rsquared_adj = 0.669), the model is improved (Table 3). Although there is a moderate correlation coefficient between MentalHealth and Poverty (Pearson's r = 0.51), MentalHealth's is still significant (P = 0.018). Because α needs to be corrected by Bofferoni to counteract the multiple comparisons problems: When choosing candidates for an additional variable, I am conducting analysis on multiple samples of data and Familywise error rate will increase.

Therefore, the corrected $\alpha = \frac{\alpha}{number\ of\ tests} = \frac{\alpha}{number\ of\ candidates} = \frac{0.1}{4} = 0.025$ ($P < \alpha$).

Elevation + Poverty (0.64)	**Mental Health**	**Asthma**	**Diabetes**	**NonWhite**
Adj R squared	**0.669**	0.657	0.647	0.65

Table 3. Adjusted R-squared for 4 candidates of the second added variable

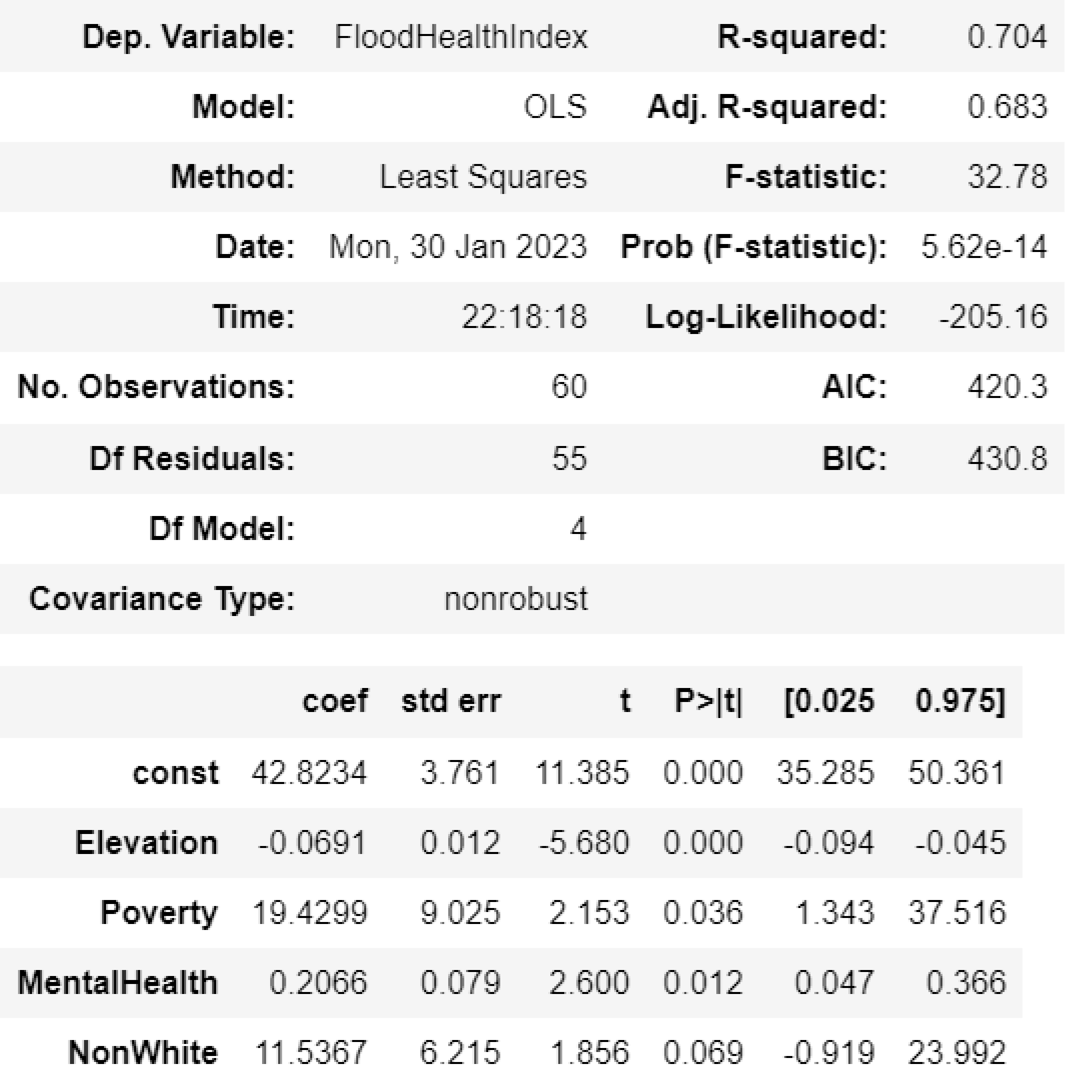
In the third round, the addition of NonWhite explains 68.3 % of the variation in Price, little more than the 66.9% explained by the previous model (Table 4).

Elevation + Poverty + MentalHealth (0.669)	**Asthma**	**Diabetes**	**NonWhite**
Adj R squared	0.664	0.664	**0.683**

Table 4. Adjusted R-squared for 3 candidates of the third added variable

However, the slope coefficient of NonWhite is not significant (P = 0.069) (Figure 5). Since P > α (corrected $\alpha = \frac{0.1}{3}$ = 0.03), NonWhite insignificantly improves the fit when elevation, poverty and mental health are included in the model. This happens possibly because NonWhite is related to Poverty (Pearson's r = 0.6): People of color often have higher poverty rates than others (Figure 6). While multicollinearity slightly increases adjusted R-squared, the significance of variable decreases. From Figure 5, Poverty is not significant anymore (p-value goes from ≈ 0 to 0.036, larger than α = 0.03), even though (by itself) Poverty has a significant linear relationship with FloodHealthIndex. Therefore, NonWhite is excluded from the model. [6]

[6] #regression: I explained why I used regression (evaluate the strength of correlation) and multiple regression (to increase prediction power). I explained the implication of slope and intercept. I explained why I prioritized R-squared over p-value in forward selection, while incorporating p-value to know when to stop adding NonWhite into the model. I kept in mind that the significance of a variable in the model only when relationship with other variables in the model (when they are held constant), by pointing out that NonWhite (alone) is not insignificant in predicting the response.



Dep. Variable:	FloodHealthIndex	**R-squared:**	0.704
Model:	OLS	**Adj. R-squared:**	0.683
Method:	Least Squares	**F-statistic:**	32.78
Date:	Mon, 30 Jan 2023	**Prob (F-statistic):**	5.62e-14
Time:	22:18:18	**Log-Likelihood:**	-205.16
No. Observations:	60	**AIC:**	420.3
Df Residuals:	55	**BIC:**	430.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	42.8234	3.761	11.385	0.000	35.285	50.361
Elevation	-0.0691	0.012	-5.680	0.000	-0.094	-0.045
Poverty	19.4299	9.025	2.153	0.036	1.343	37.516
MentalHealth	0.2066	0.079	2.600	0.012	0.047	0.366
NonWhite	11.5367	6.215	1.856	0.069	-0.919	23.992

Figure 5. The summary of the regression model with Elevation, Poverty, Mental Health, and NonWhite

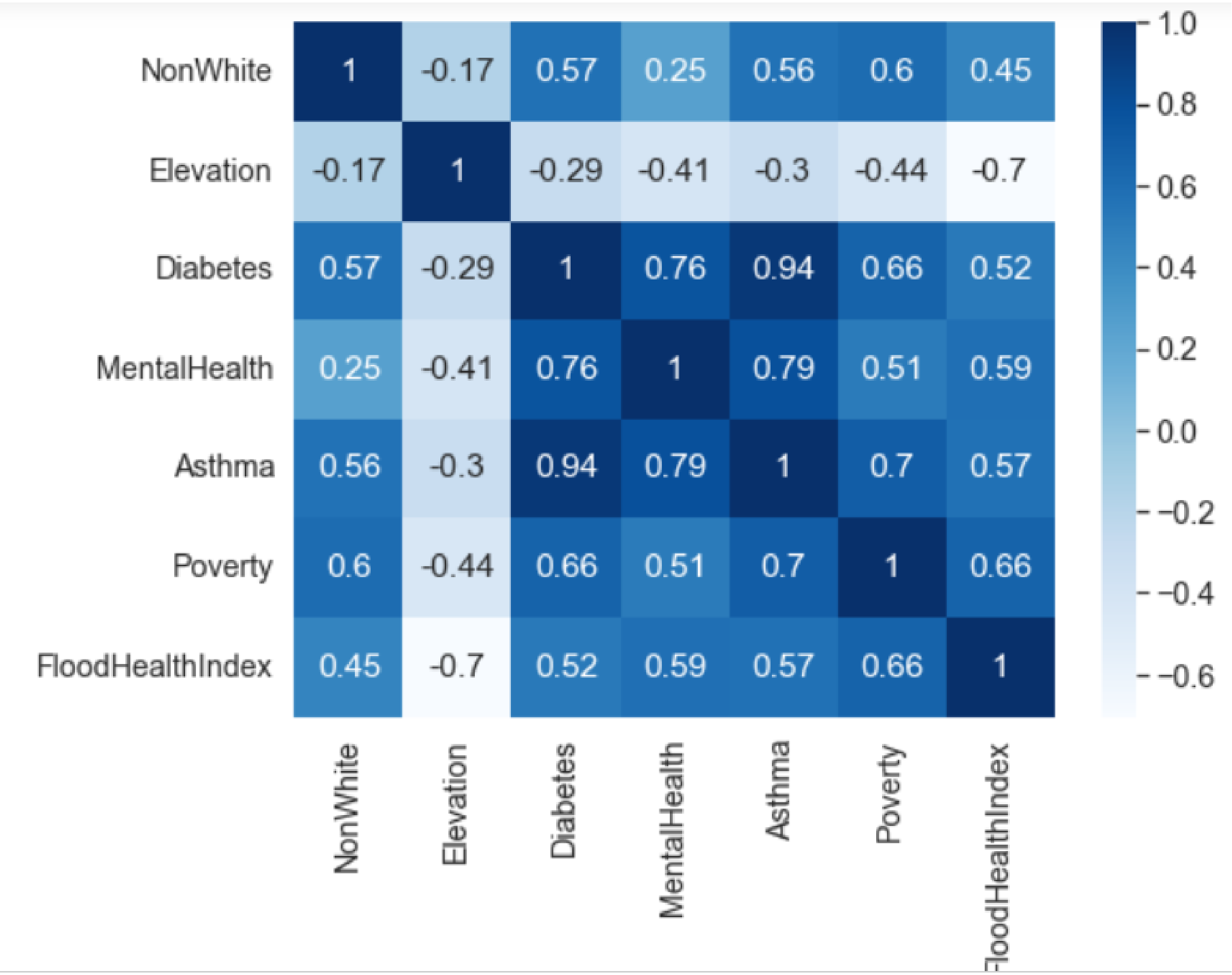


Figure 6. Correlation matrix of predictor variables after removing non-linear variables.

Multiple regression equation

FloodHealthIndex = -0.067 * Elevation + 29.052 * Poverty + 0.198 * MentalHealth + 46.715. An increase of Elevation by 1 foot decreases FloodHealthIndex by 0.067 points, provided other variables are held constant. An increase of Poverty by 1 point percentage of all individuals below 200% of the Federal poverty rate increases FloodHealthIndex by 29.052 points, provided other variables are held constant. An increase of MentalHealth by 1 age-adjusted hospitalization rate per 100,000 residents due to schizophrenia and other psychotic disorders increases FloodHealthIndex by 0.198 points, provided other variables are held constant. When all predictor variables are equal to 0, which is not practical and leads to extrapolation, FloodHealthIndex will be 46.715 points.

Conclusion

In single regression, there is a strong and negative linear relationship between elevation and FloodHealthIndex. The slope coefficient of elevation is statistically significant, and elevation explains 49.5% of the variation in the FloodHealthIndex. We can be 90% confident that each additional foot of the minimum elevation would expectedly decrease the FloodHealthIndex somewhere between 0.08 and 0.123 points [7]. The adjusted R-squared of *the multiple regression* increases to 66.9% indicating a better fit to the data, while each variable is still significant.

Applications: Policymakers could predict FloodHealthIndex of an area based on inputs of the local elevation level, poverty, and mental health rate; thus predicting resource allocation. NonWhite level does not help predict FloodHealthIndex, given that other variables are present. This does not mean that NonWhite alone cannot help predict FloodHealthIndex, but exclusion is needed for the model to be simple, cost-effective, and avoid overfitting and multicollinearity. One practical limitation is that while it makes sense for an area's elevation to remain fixed, it is impractical to keep poverty and mental health fixed.

Inference: Of the population of 578 census blockgroups, the unknown population slope coefficient β1 in simple regression is now estimated by the point estimates of the slope *b1* to be between -0.123 and -0.08, captured by 90% of the times conducting confidence intervals. This is a generalization induction made from a Frequentist approach because given all true premises about R-squared, significant p-value, confidence interval, and the fact that conditions are met and

[7] #probability: I explained p-value, significance test, confidence interval, and R-squared in terms of probability. I applied knowledge of the Frequentist approach in interpreting those values, for example, when sampling infinitely in the population, what would be the probability that I obtained the observed fitted data points of response variable that can be explained by the model, or the probability that the confidence interval can capture the true population slope coefficient.

we can sample infinitely from the population, the above conclusions are likely to be true. The small p-value and narrow confidence interval strengthen the induction because they increase the probability of rejecting the null and decrease the probability of having a Type II error. However, reliability is not ensured. The inference conditions - Independence is not guaranteed since the sample is smaller than 10% of the population. There are unequal variances of residuals, we cannot infer the true shape of the population's scatterplot because due to the non-football shape, few datapoints are on the right-hand side. Therefore, population inference needs to be cautious with extrapolation since the observed linear relationship is limited to the given range of sample data. Therefore, violations of both in the population (dependence and heteroscedasticity) would break the premises because the standard error or distribution assumption of the point estimate – assumed to be normal when applying the t-test statistic – may not be valid, which in turn falsifies premises of p-value or confidence interval. Extraneous variables could also break the premises, for example, an increase in Hospitalization Cost possibly increases the moderate correlation between Poverty and Mental Health (Pearson's r = 0.51), thus p-value and making the model insignificant. Lastly, the model cannot infer any causation, further tests are needed. [8] [9]

Word count: 1800 words

[8] # organization: : I stated the research question, the objective and the outline at the introduction. The conclusion directly answers the research question. I applied the figures and tables when they are most important. For example, the exhaustive scatter plots between 17 independent variables and dependent variable have been shorten when I provided 4 scatter plots that shows non-linearity. The shorten list of variables helps increase effectiveness of forward selection.

[9] # induction: I specified what is the premises and the conclusion. I examined the strength using direct result such as how the value of p-value and confidence interval affect the strength. The reliability was examined using the evaluation of conditions and provided limitation. I used examples when the reliability should be doubt, such as when the data range of the population is not reflected in the sample data, when making extrapolation, and when there are Extraneous variable and warning about causation.

Reflection [10]

1. Explain one method you used to determine if your results are correct. For example, how do you know that the t-scores for the confidence intervals or significance test are accurate? How do you know that your final confidence interval and/or p-value are correct? (<75 words)

I would compare the consistency between results and conduct multiple calculations of a metric with different formulas or in different ways (python, online calculator…). For example, I double-checked the p-value and confidence interval using codes by hand instead of merely relying on the summary table of the regression model. I looked up the internet to see in detail the value that the regression model output and compare with what I need.

2. Most scientific research papers end with acknowledgments. For this assignment, give acknowledgment to the people and/or external resources that you used to complete it. This doesn't have to be an exhaustive list, but is a chance for an honest reflection and recognition (<75 words).

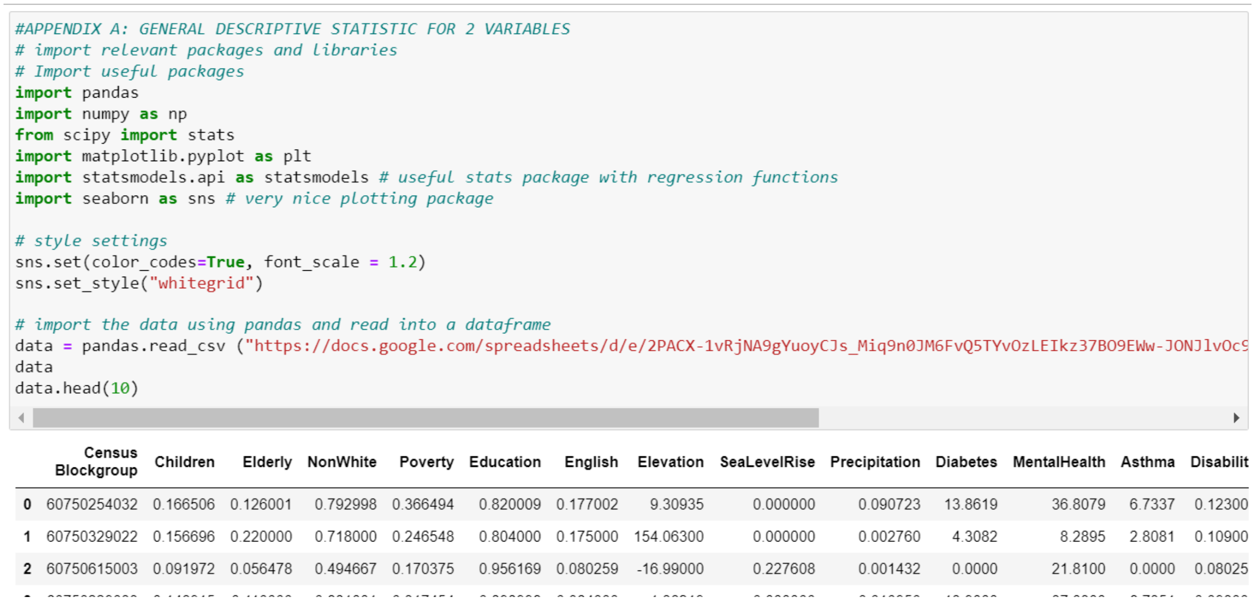
I apply knowledge from the OpenStats textbook and multiple in-class activities (algorithms, codes…). Moreover, special thanks to Rawan for correcting my coding mistakes in constructing the confidence interval and p-value (I forgot to output the absolute value of T-score so that leads to a false p-value).

[10] #professionalism: I structured the report and the appendix well. I followed a 2-round checking of grammar mistakes, cohesion and logic. I also sited relevant resources that I used in the report.

APPENDIX

APPENDIX A: IMPORT, GENERAL DATASET AND VARIABLE INFORMATION

(1) Dataset import and display



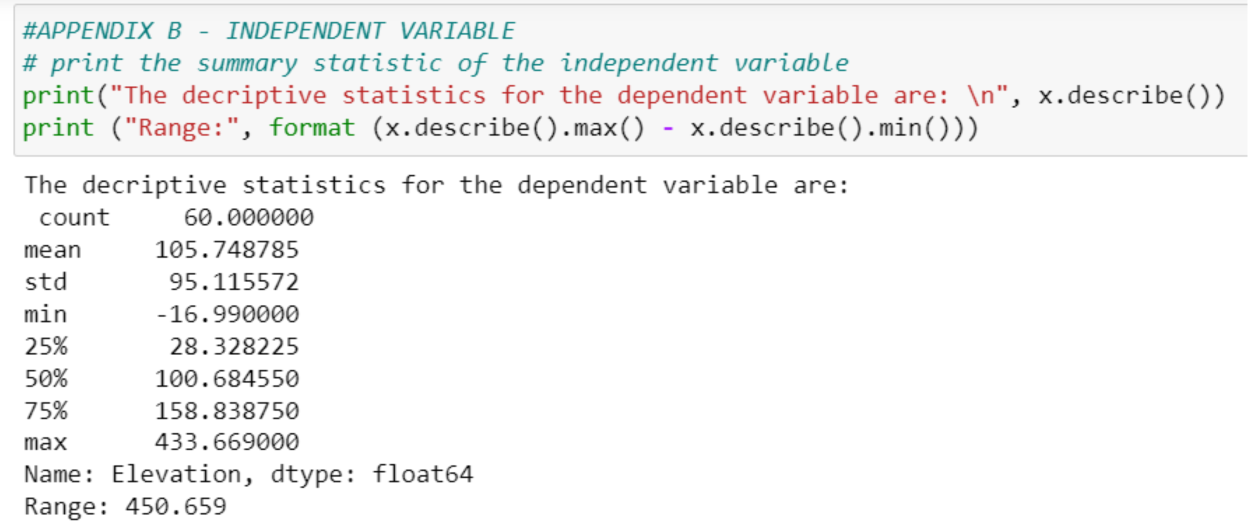
```
#APPENDIX A: GENERAL DESCRIPTIVE STATISTIC FOR 2 VARIABLES
# import relevant packages and libraries
# Import useful packages
import pandas
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import statsmodels.api as statsmodels # useful stats package with regression functions
import seaborn as sns # very nice plotting package

# style settings
sns.set(color_codes=True, font_scale = 1.2)
sns.set_style("whitegrid")

# import the data using pandas and read into a dataframe
data = pandas.read_csv ("https://docs.google.com/spreadsheets/d/e/2PACX-1vRjNA9gYuoyCJs_Miq9n0JM6FvQ5TYvOzLEIkz37BO9EWw-JONJlvOc9
data
data.head(10)
```

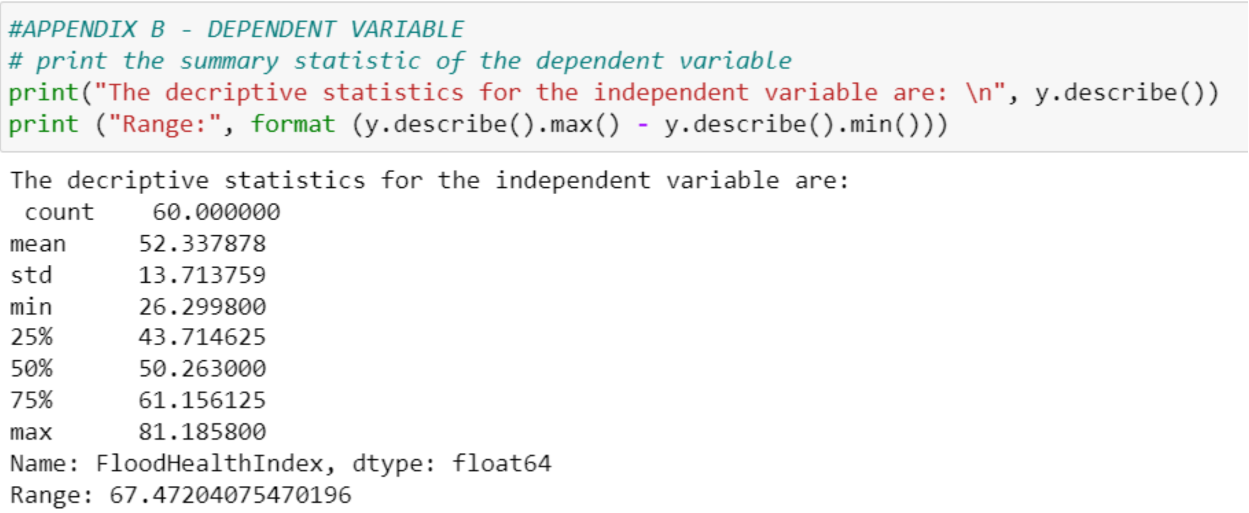
	Census Blockgroup	Children	Elderly	NonWhite	Poverty	Education	English	Elevation	SeaLevelRise	Precipitation	Diabetes	MentalHealth	Asthma	Disabilit
0	60750254032	0.166506	0.126001	0.792998	0.366494	0.820009	0.177002	9.30935	0.000000	0.090723	13.8619	36.8079	6.7337	0.12300
1	60750329022	0.156696	0.220000	0.718000	0.246548	0.804000	0.175000	154.06300	0.000000	0.002760	4.3082	8.2895	2.8081	0.10900
2	60750615003	0.091972	0.056478	0.494667	0.170375	0.956169	0.080259	-16.99000	0.227608	0.001432	0.0000	21.8100	0.0000	0.08025

(1) General descriptive statistics and data visualization for 2 variables



```
#APPENDIX B - INDEPENDENT VARIABLE
# print the summary statistic of the independent variable
print("The decriptive statistics for the dependent variable are: \n", x.describe())
print ("Range:", format (x.describe().max() - x.describe().min()))
```

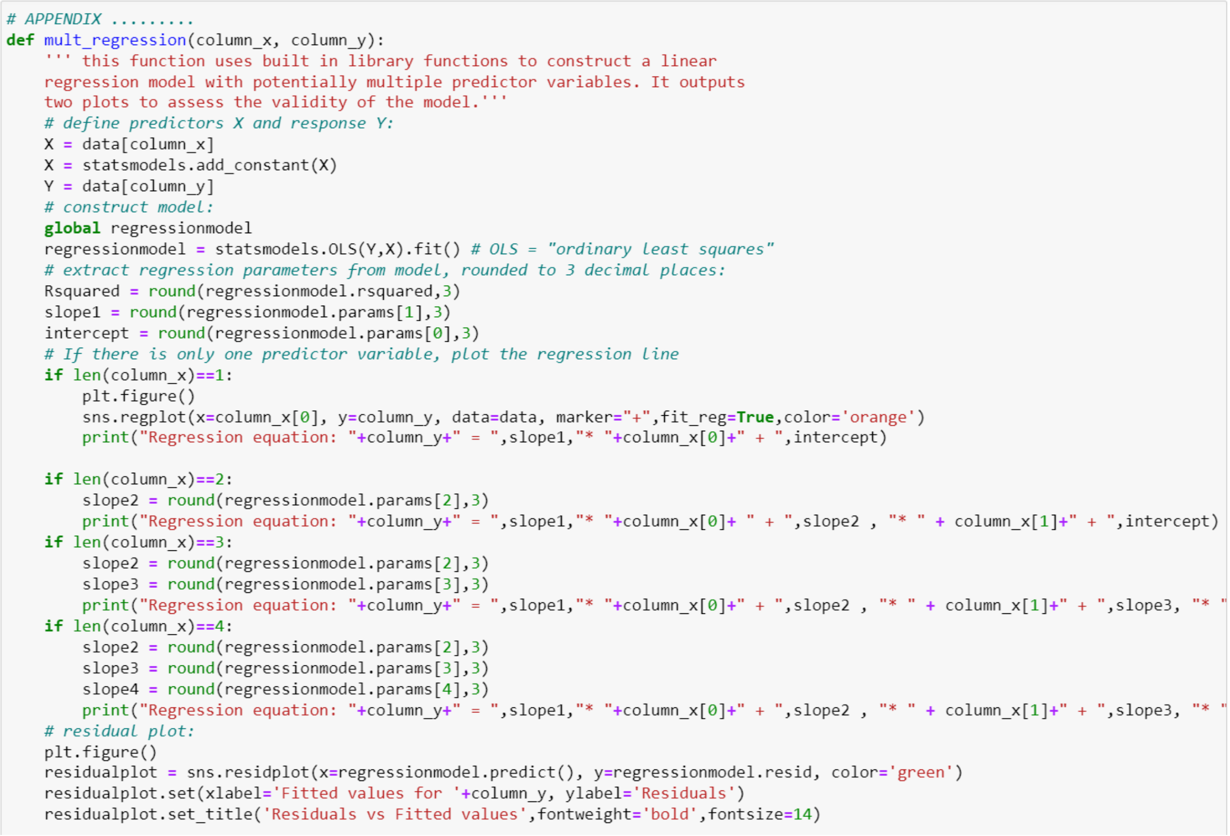
```
The decriptive statistics for the dependent variable are:
 count      60.000000
mean      105.748785
std        95.115572
min       -16.990000
25%        28.328225
50%       100.684550
75%       158.838750
max       433.669000
Name: Elevation, dtype: float64
Range: 450.659
```



```
#APPENDIX B - DEPENDENT VARIABLE
# print the summary statistic of the dependent variable
print("The decriptive statistics for the independent variable are: \n", y.describe())
print ("Range:", format (y.describe().max() - y.describe().min()))
```

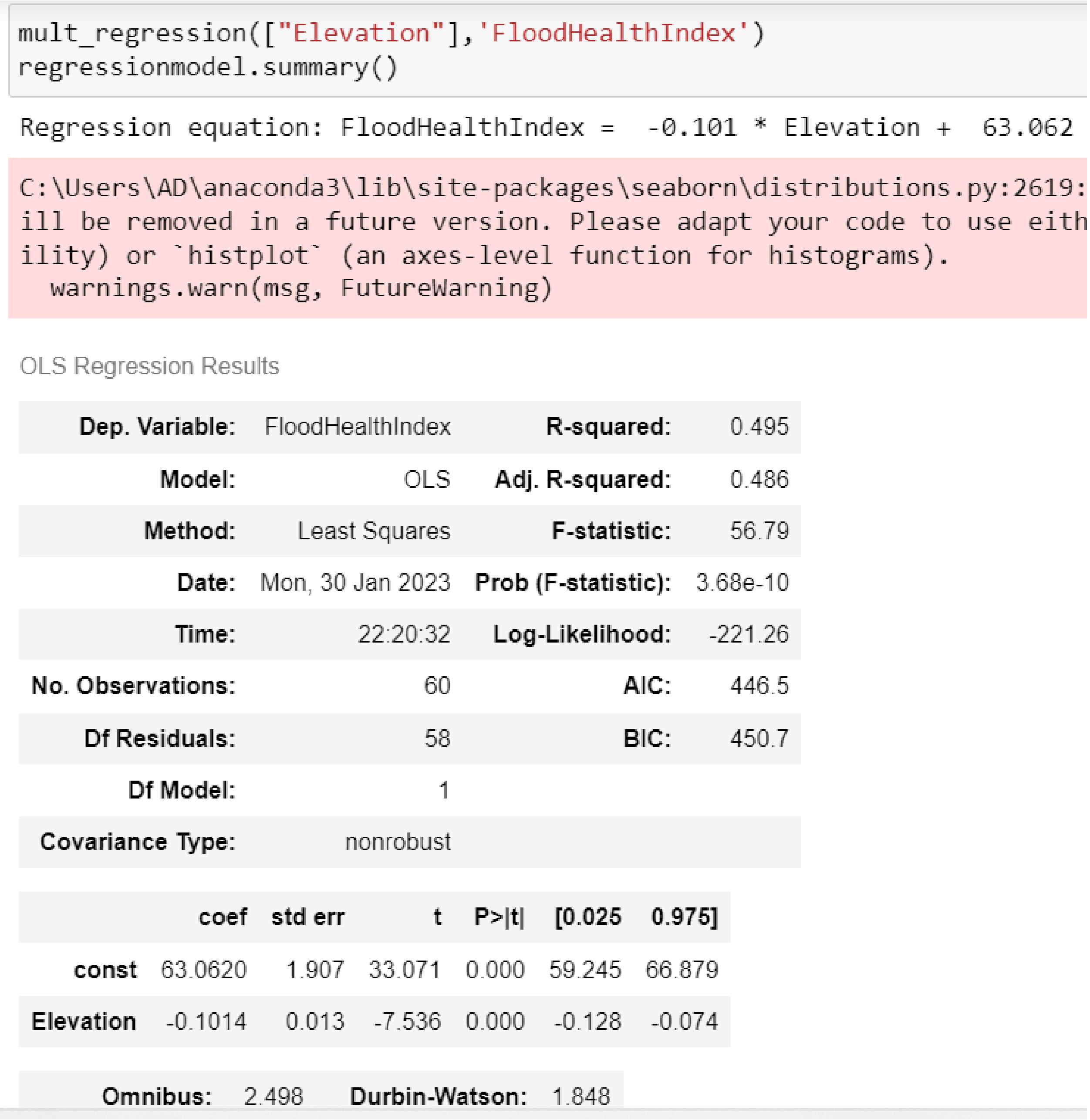
```
The decriptive statistics for the independent variable are:
 count    60.000000
mean     52.337878
std      13.713759
min      26.299800
25%      43.714625
50%      50.263000
75%      61.156125
max      81.185800
Name: FloodHealthIndex, dtype: float64
Range: 67.47204075470196
```

APPENDIX C:



```
# APPENDIX .........
def mult_regression(column_x, column_y):
    ''' this function uses built in library functions to construct a linear
    regression model with potentially multiple predictor variables. It outputs
    two plots to assess the validity of the model.'''
    # define predictors X and response Y:
    X = data[column_x]
    X = statsmodels.add_constant(X)
    Y = data[column_y]
    # construct model:
    global regressionmodel
    regressionmodel = statsmodels.OLS(Y,X).fit() # OLS = "ordinary least squares"
    # extract regression parameters from model, rounded to 3 decimal places:
    Rsquared = round(regressionmodel.rsquared,3)
    slope1 = round(regressionmodel.params[1],3)
    intercept = round(regressionmodel.params[0],3)
    # If there is only one predictor variable, plot the regression line
    if len(column_x)==1:
        plt.figure()
        sns.regplot(x=column_x[0], y=column_y, data=data, marker="+",fit_reg=True,color='orange')
        print("Regression equation: "+column_y+" = ",slope1,"* "+column_x[0]+" + ",intercept)

    if len(column_x)==2:
        slope2 = round(regressionmodel.params[2],3)
        print("Regression equation: "+column_y+" = ",slope1,"* "+column_x[0]+ " + ",slope2 , "* " + column_x[1]+" + ",intercept)
    if len(column_x)==3:
        slope2 = round(regressionmodel.params[2],3)
        slope3 = round(regressionmodel.params[3],3)
        print("Regression equation: "+column_y+" = ",slope1,"* "+column_x[0]+" + ",slope2 , "* " + column_x[1]+" + ",slope3, "* "
    if len(column_x)==4:
        slope2 = round(regressionmodel.params[2],3)
        slope3 = round(regressionmodel.params[3],3)
        slope4 = round(regressionmodel.params[4],3)
        print("Regression equation: "+column_y+" = ",slope1,"* "+column_x[0]+" + ",slope2 , "* " + column_x[1]+" + ",slope3, "* "
    # residual plot:
    plt.figure()
    residualplot = sns.residplot(x=regressionmodel.predict(), y=regressionmodel.resid, color='green')
    residualplot.set(xlabel='Fitted values for '+column_y, ylabel='Residuals')
    residualplot.set_title('Residuals vs Fitted values',fontweight='bold',fontsize=14)
```



```
mult_regression(["Elevation"],'FloodHealthIndex')
regressionmodel.summary()
```

```
Regression equation: FloodHealthIndex =  -0.101 * Elevation +  63.062
```

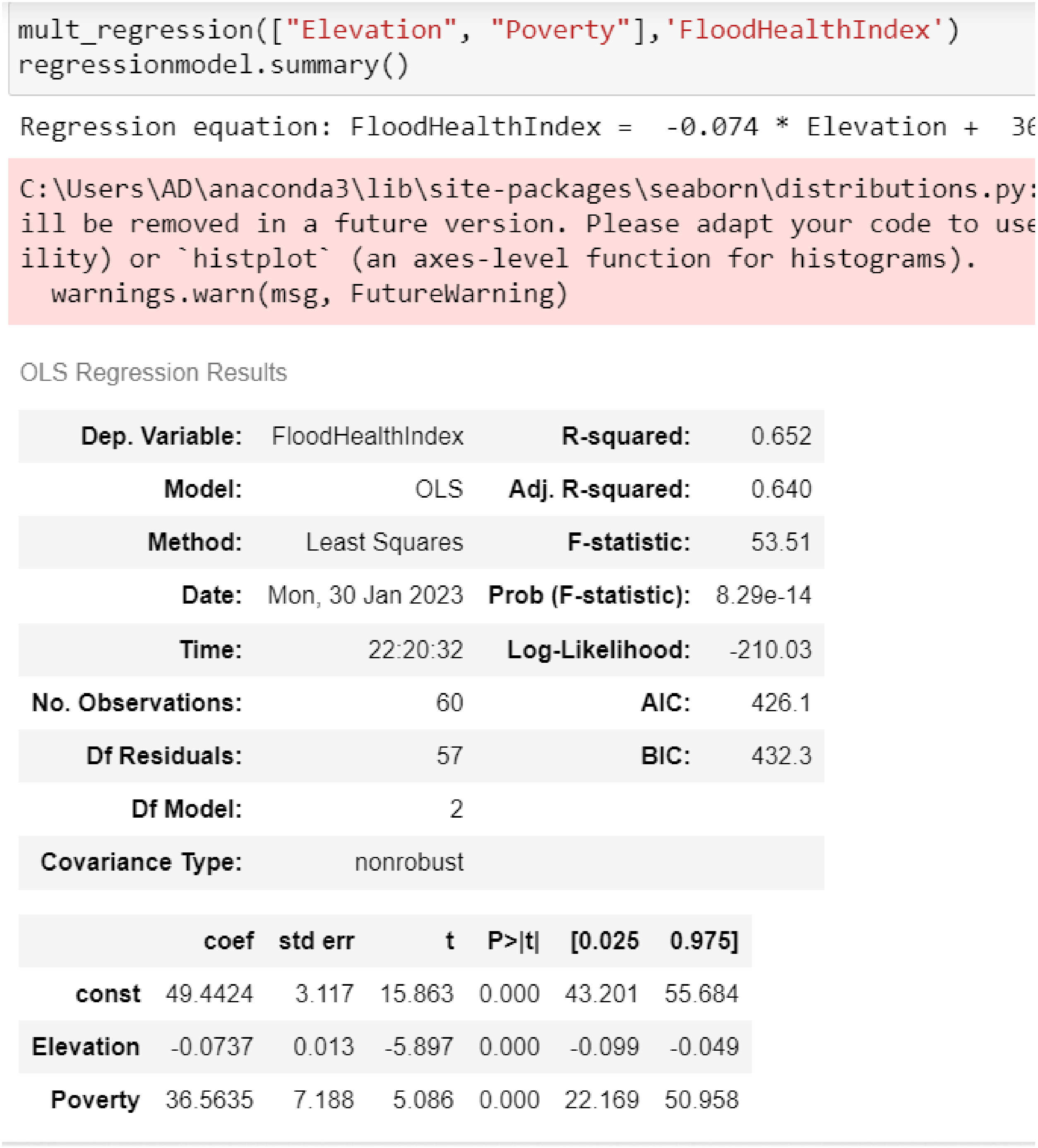
```
C:\Users\AD\anaconda3\lib\site-packages\seaborn\distributions.py:2619:
ill be removed in a future version. Please adapt your code to use eith
ility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

OLS Regression Results

Dep. Variable:	FloodHealthIndex	**R-squared:**	0.495
Model:	OLS	**Adj. R-squared:**	0.486
Method:	Least Squares	**F-statistic:**	56.79
Date:	Mon, 30 Jan 2023	**Prob (F-statistic):**	3.68e-10
Time:	22:20:32	**Log-Likelihood:**	-221.26
No. Observations:	60	**AIC:**	446.5
Df Residuals:	58	**BIC:**	450.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	63.0620	1.907	33.071	0.000	59.245	66.879
Elevation	-0.1014	0.013	-7.536	0.000	-0.128	-0.074

Omnibus:	2.498	**Durbin-Watson:**	1.848



```
mult_regression(["Elevation", "Poverty"],'FloodHealthIndex')
regressionmodel.summary()
```

```
Regression equation: FloodHealthIndex =  -0.074 * Elevation +  36
```

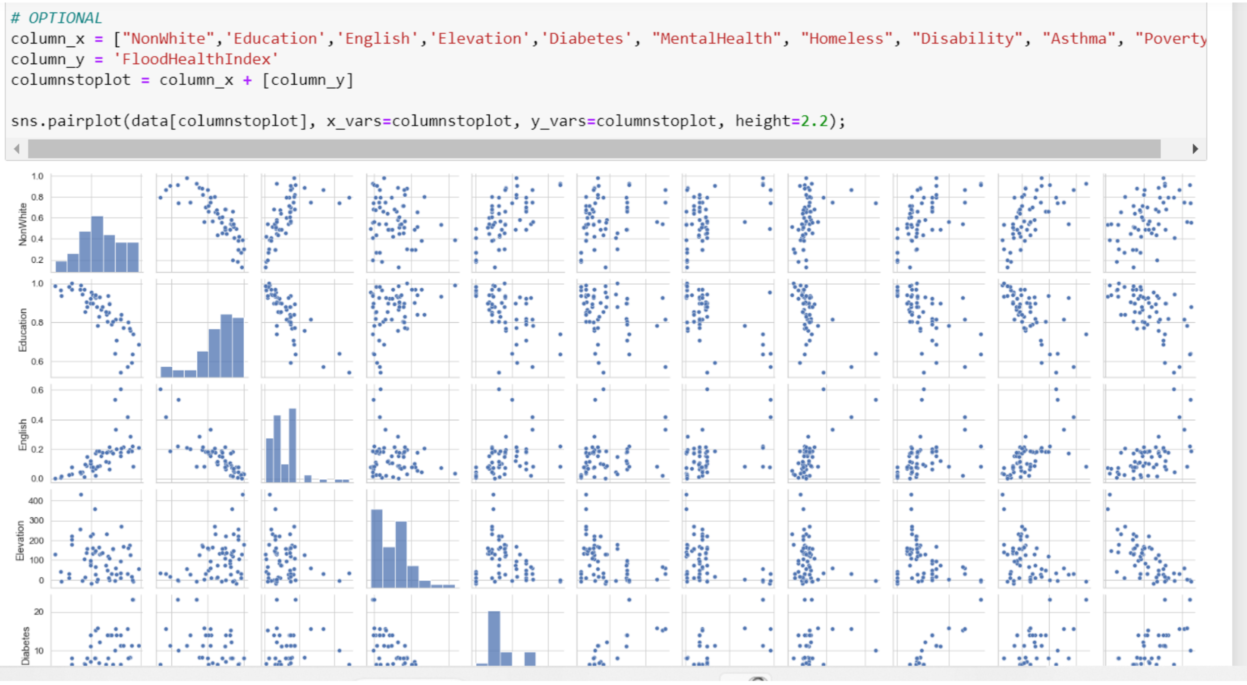
```
C:\Users\AD\anaconda3\lib\site-packages\seaborn\distributions.py:
ill be removed in a future version. Please adapt your code to use
ility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

OLS Regression Results

Dep. Variable:	FloodHealthIndex	**R-squared:**	0.652
Model:	OLS	**Adj. R-squared:**	0.640
Method:	Least Squares	**F-statistic:**	53.51
Date:	Mon, 30 Jan 2023	**Prob (F-statistic):**	8.29e-14
Time:	22:20:32	**Log-Likelihood:**	-210.03
No. Observations:	60	**AIC:**	426.1
Df Residuals:	57	**BIC:**	432.3
Df Model:	2		
Covariance Type:	nonrobust		

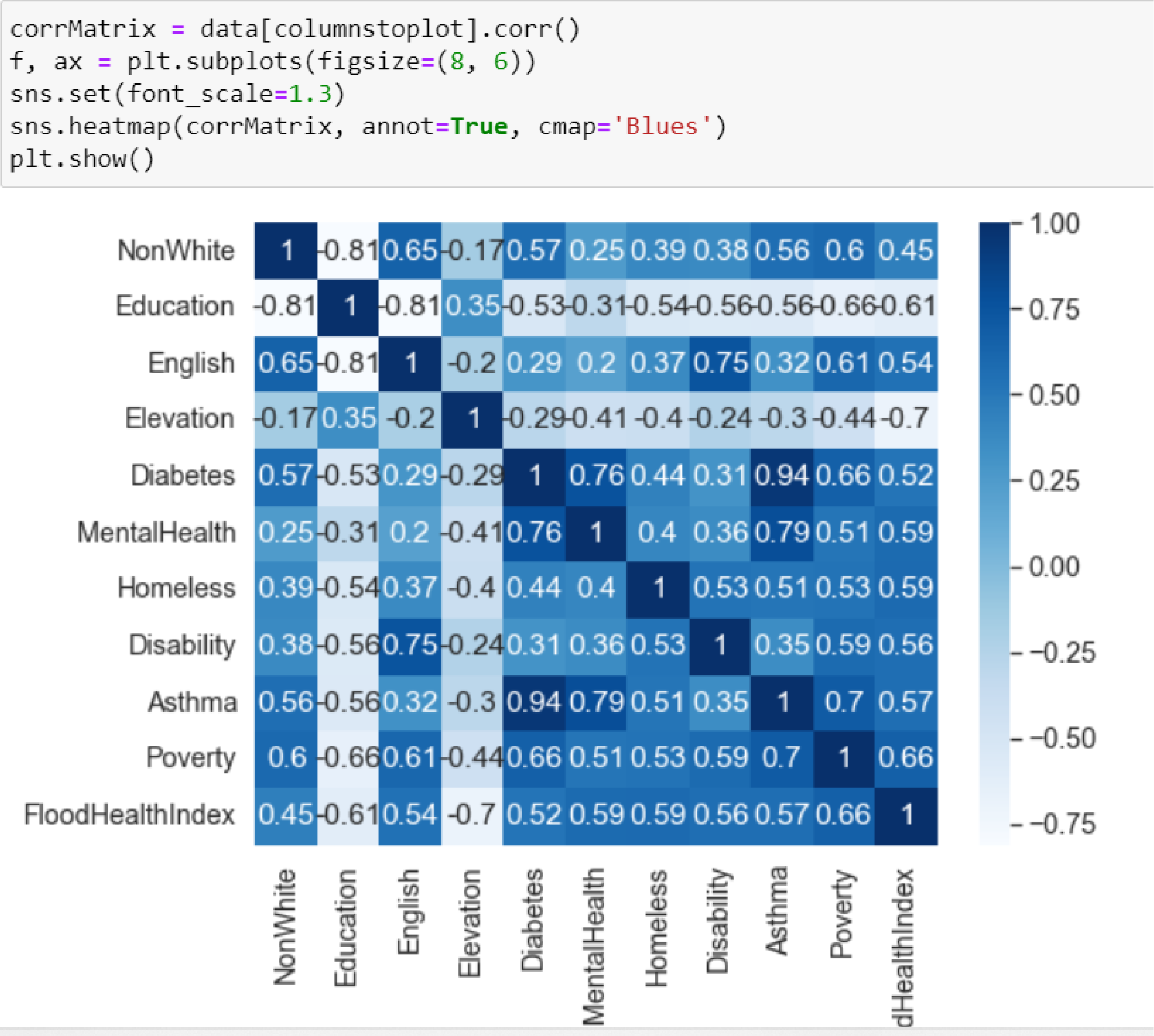
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	49.4424	3.117	15.863	0.000	43.201	55.684
Elevation	-0.0737	0.013	-5.897	0.000	-0.099	-0.049
Poverty	36.5635	7.188	5.086	0.000	22.169	50.958

APPENDIX D:



```
# OPTIONAL
column_x = ["NonWhite",'Education','English','Elevation','Diabetes', "MentalHealth", "Homeless", "Disability", "Asthma", "Poverty
column_y = 'FloodHealthIndex'
columnstoplot = column_x + [column_y]

sns.pairplot(data[columnstoplot], x_vars=columnstoplot, y_vars=columnstoplot, height=2.2);
```



```
corrMatrix = data[columnstoplot].corr()
f, ax = plt.subplots(figsize=(8, 6))
sns.set(font_scale=1.3)
sns.heatmap(corrMatrix, annot=True, cmap='Blues')
plt.show()
```

	NonWhite	Education	English	Elevation	Diabetes	MentalHealth	Homeless	Disability	Asthma	Poverty	FloodHealthIndex
NonWhite	1	-0.81	0.65	-0.17	0.57	0.25	0.39	0.38	0.56	0.6	0.45
Education	-0.81	1	-0.81	0.35	-0.53	-0.31	-0.54	-0.56	-0.56	-0.66	-0.61
English	0.65	-0.81	1	-0.2	0.29	0.2	0.37	0.75	0.32	0.61	0.54
Elevation	-0.17	0.35	-0.2	1	-0.29	-0.41	-0.4	-0.24	-0.3	-0.44	-0.7
Diabetes	0.57	-0.53	0.29	-0.29	1	0.76	0.44	0.31	0.94	0.66	0.52
MentalHealth	0.25	-0.31	0.2	-0.41	0.76	1	0.4	0.36	0.79	0.51	0.59
Homeless	0.39	-0.54	0.37	-0.4	0.44	0.4	1	0.53	0.51	0.53	0.59
Disability	0.38	-0.56	0.75	-0.24	0.31	0.36	0.53	1	0.35	0.59	0.56
Asthma	0.56	-0.56	0.32	-0.3	0.94	0.79	0.51	0.35	1	0.7	0.57
Poverty	0.6	-0.66	0.61	-0.44	0.66	0.51	0.53	0.59	0.7	1	0.66
FloodHealthIndex	0.45	-0.61	0.54	-0.7	0.52	0.59	0.59	0.56	0.57	0.66	1