Raquel Ocasio

D208 – Predictive Modeling, Task 1

October 3, 2023

Western Governors University

## Part A1

Which variables influence the number of days the patient stayed in the hospital during the initial visit?

## Part A2

The goal of the data analysis is to determine if the number of days the patient stayed in the hospital is influenced by other variables in the dataset.

## Part B1

Multiple linear regression assumes that there is a linear relationship between the dependent variable and a set of independent variables. It assumes that the errors (residuals) are independently and identically distributed, meaning they have constant variance and are uncorrelated. Homoscedasticity is another key assumption, suggesting that the variance of the residuals remains consistent across all levels of the independent variables. Finally, there should be minimal multicollinearity among the independent variables, meaning they are not highly correlated with each other, as high multicollinearity can make it challenging to discern the unique effects of individual predictors on the dependent variable. (GeeksforGeeks, 2023)

## Part B2

Two benefits of using Python for the analysis are the libraries, and access to additional support. First, Python provides a large selection of libraries and tools that simplify the process of building, training, and evaluating multiple linear regression models. Second, Python is open-source with a large and active community, ensuring access to a wealth of resources, tutorials, and community support.

## Part B3

The target variable for this analysis is continuous. Multiple linear regression is appropriate for this analysis because it can model the relationship between a continuous response variable and one or more explanatory variables that are continuous and/or categorical.

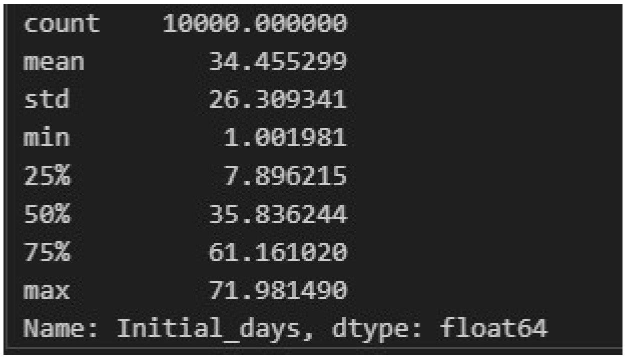
## Part C1

The goals of the data cleaning process are to detect and treat duplicate values, missing values, and outlier values. Duplicate values are detected using the .duplicated().value_counts() functions. No duplicate values were detected. Missing values are detected using the .isnull().sum() functions. No missing values were detected. The detection of outliers is a three-step process. First, the z-scores are calculated using the stats.zscore() function. Second, the values with a z-score of less than -3 or greater than 3 are saved to a new dataframe using the .query() function. Third, the number of observations in the new dataframe is counted using the len() function. Outliers were detected and retained.

See attached code.

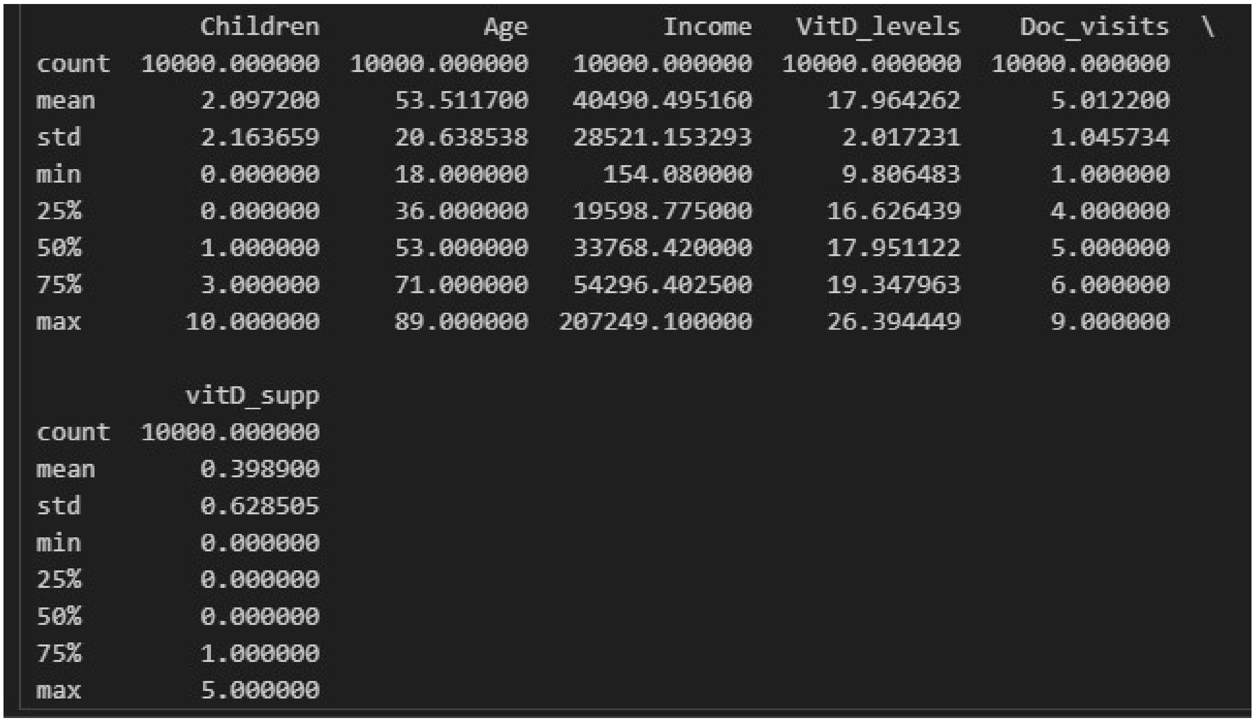
## Part C2

Summary statistics for dependent variable (Verma, 2020)



```
count    10000.000000
mean        34.455299
std         26.309341
min          1.001981
25%          7.896215
50%         35.836244
75%         61.161020
max         71.981490
Name: Initial_days, dtype: float64
```

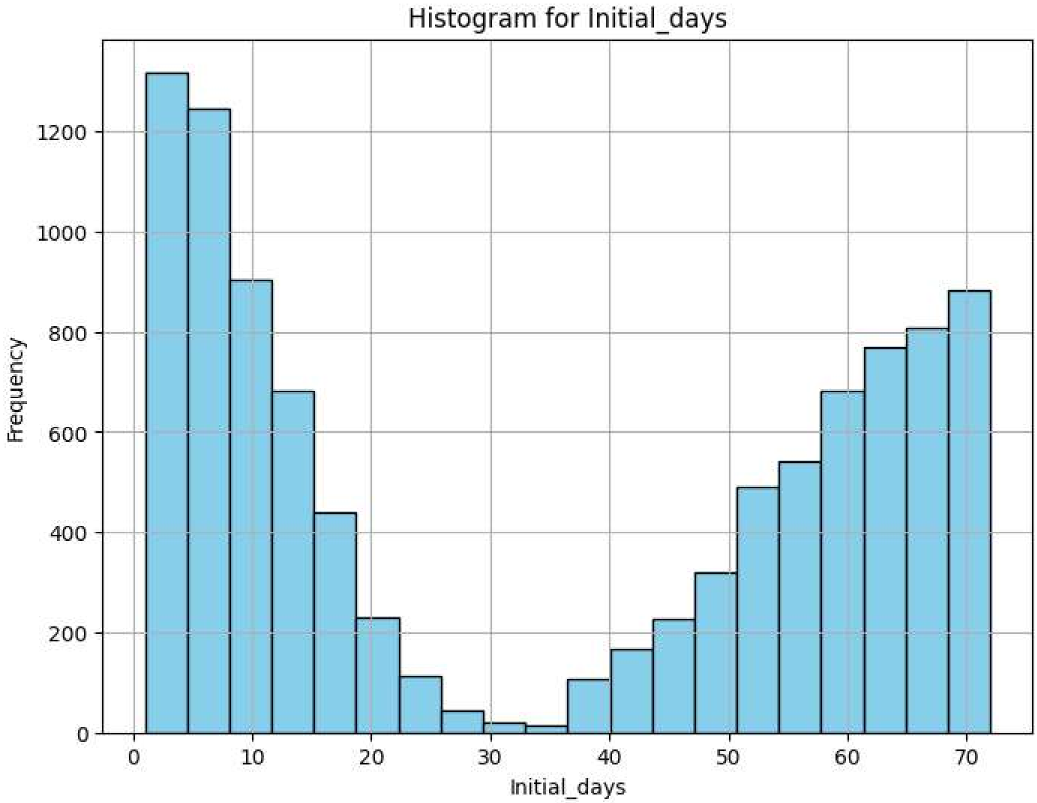
Summary statistics for independent variables (Verma, 2020)

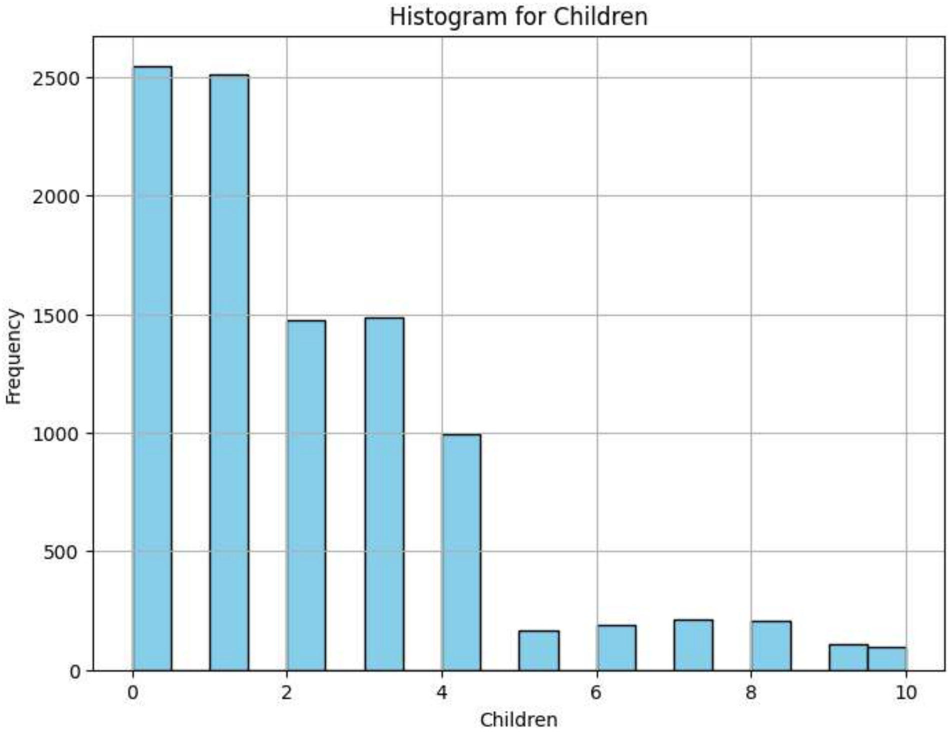


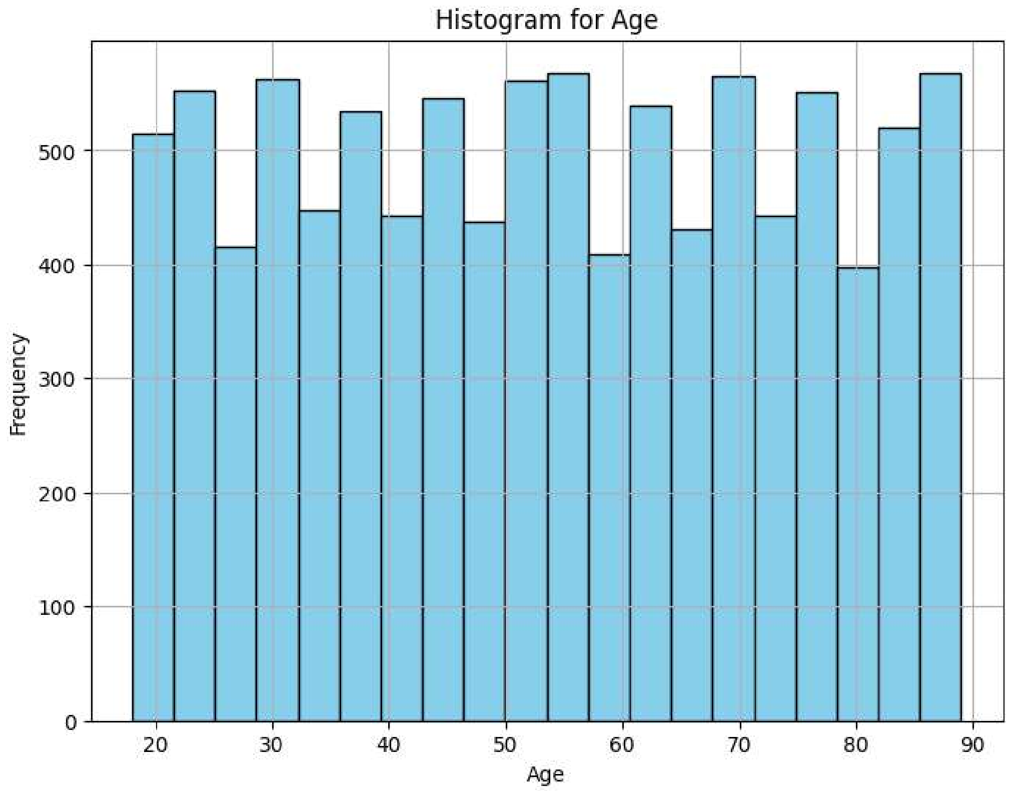
```
           Children           Age         Income   VitD_levels    Doc_visits  \
count  10000.000000  10000.000000   10000.000000  10000.000000  10000.000000
mean       2.097200     53.511700   40490.495160     17.964262      5.012200
std        2.163659     20.638538   28521.153293      2.017231      1.045734
min        0.000000     18.000000     154.080000      9.806483      1.000000
25%        0.000000     36.000000   19598.775000     16.626439      4.000000
50%        1.000000     53.000000   33768.420000     17.951122      5.000000
75%        3.000000     71.000000   54296.402500     19.347963      6.000000
max       10.000000     89.000000  207249.100000     26.394449      9.000000

          vitD_supp
count  10000.000000
mean       0.398900
std        0.628505
min        0.000000
25%        0.000000
50%        0.000000
75%        1.000000
max        5.000000
```

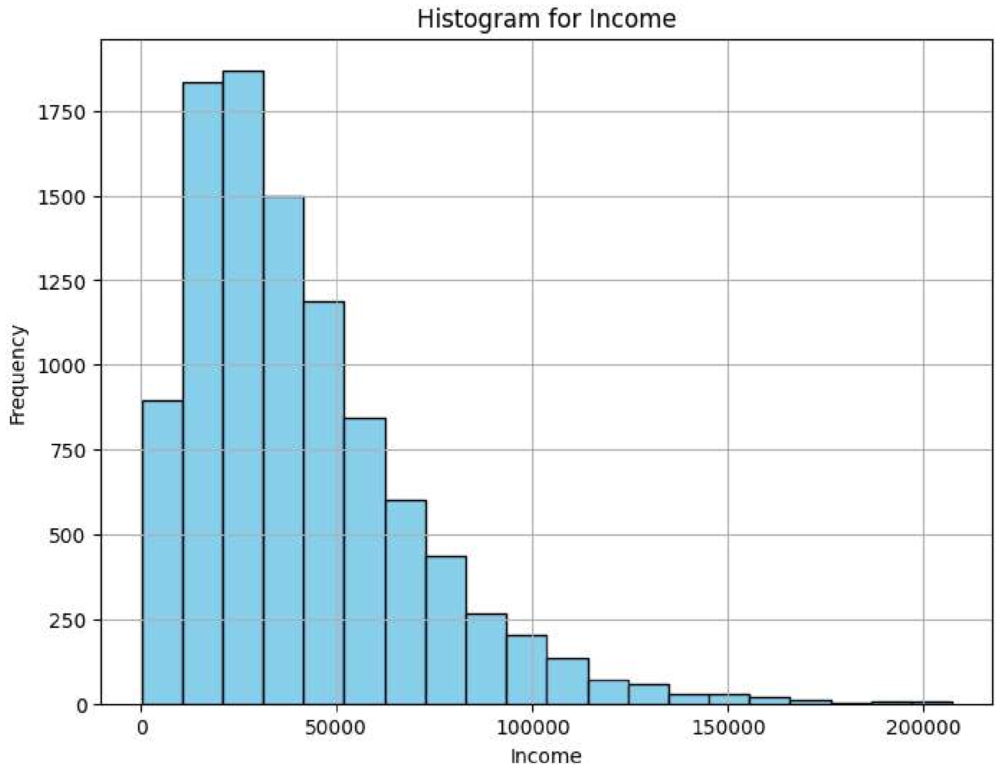
## Part C3

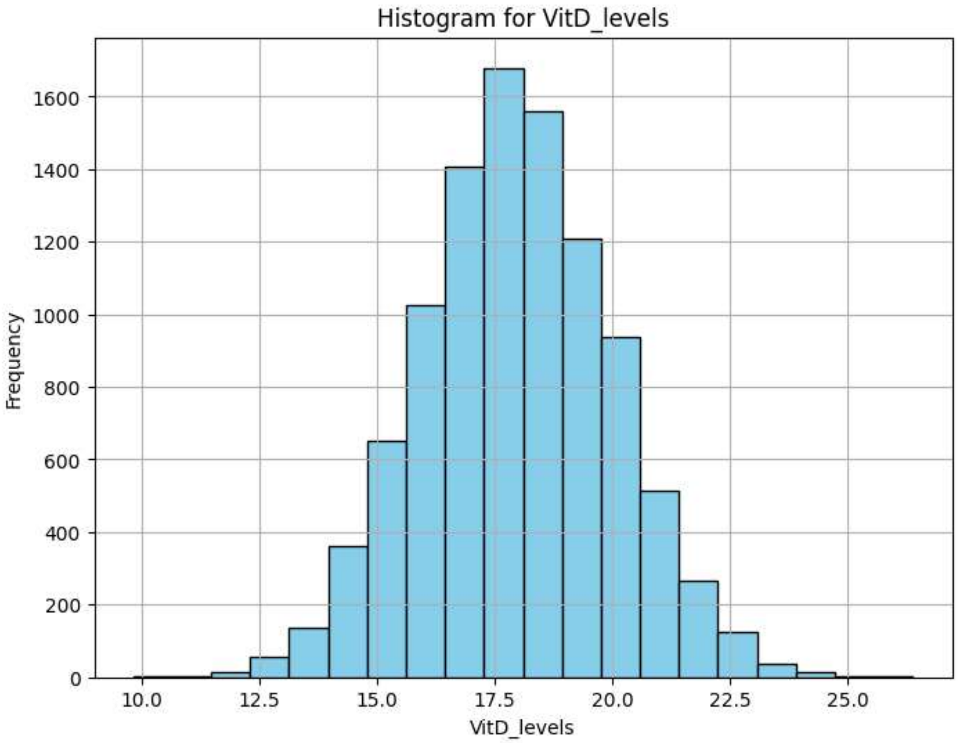
Univariate visualizations

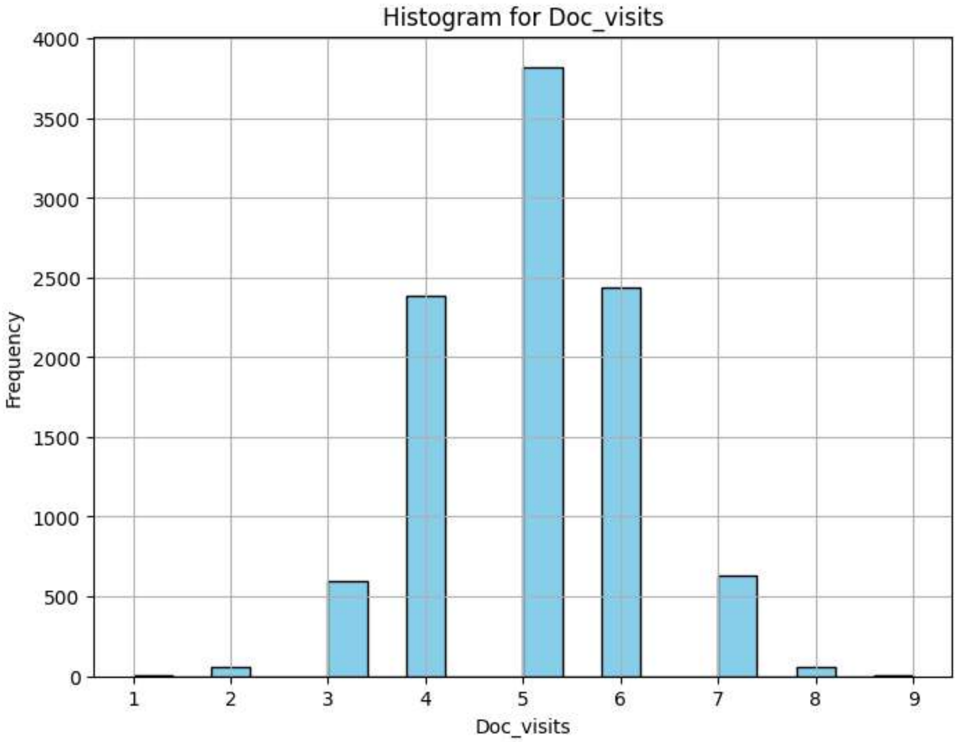


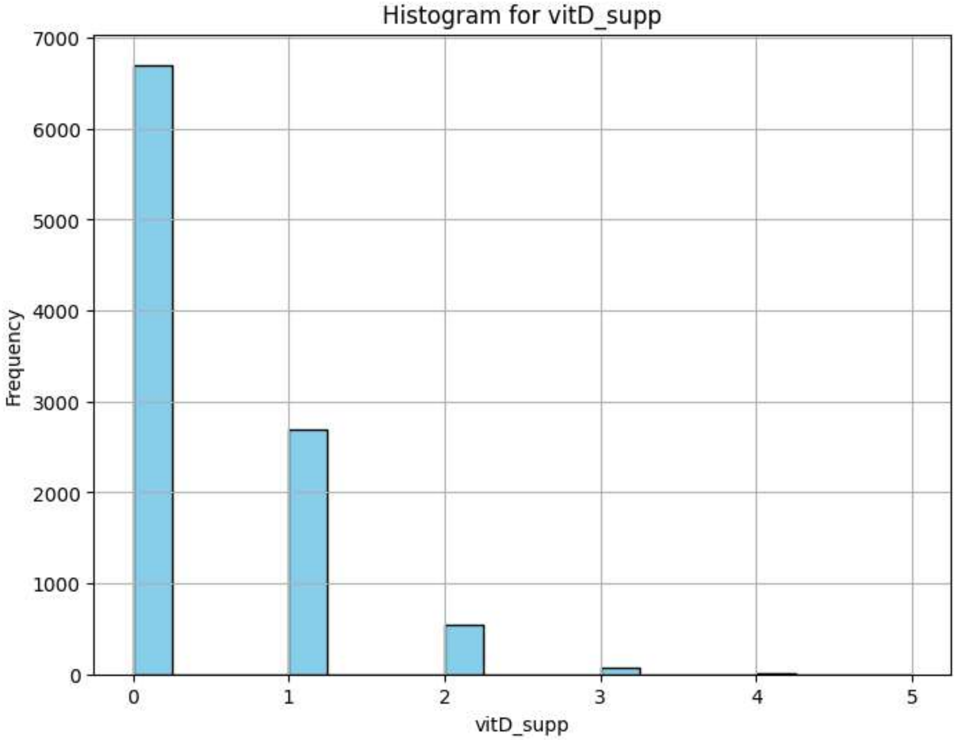



Histogram for Age
Frequency
0
100
200
300
400
500
20
30
40
50
60
70
80
90
Age

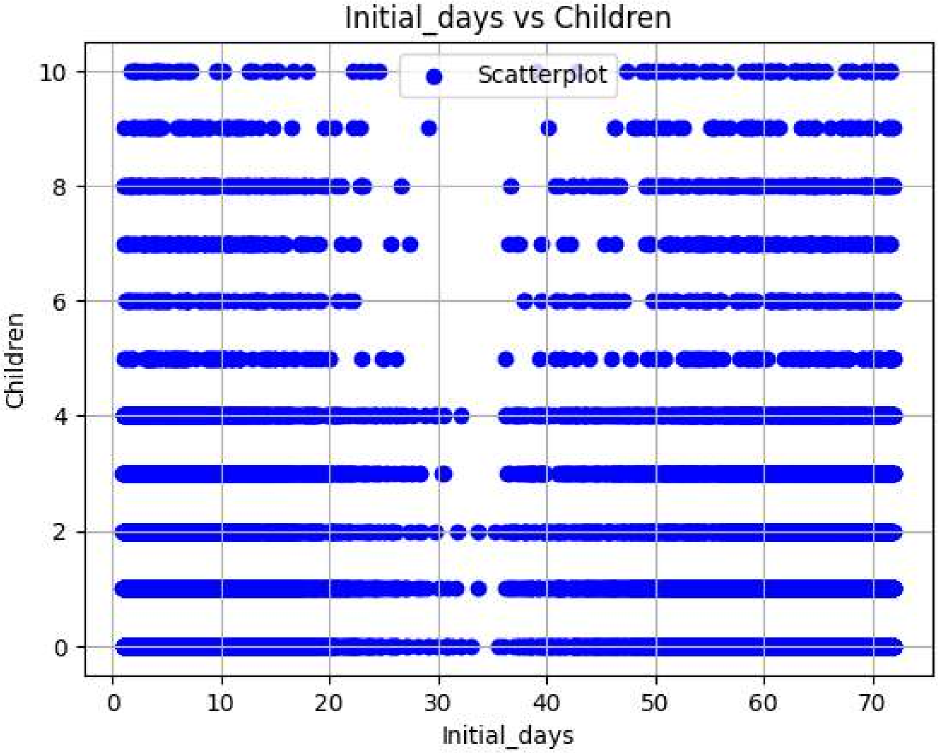

Histogram for Income
Frequency
0
250
500
750
1000
1250
1500
1750
0
50000
100000
150000
200000
Income

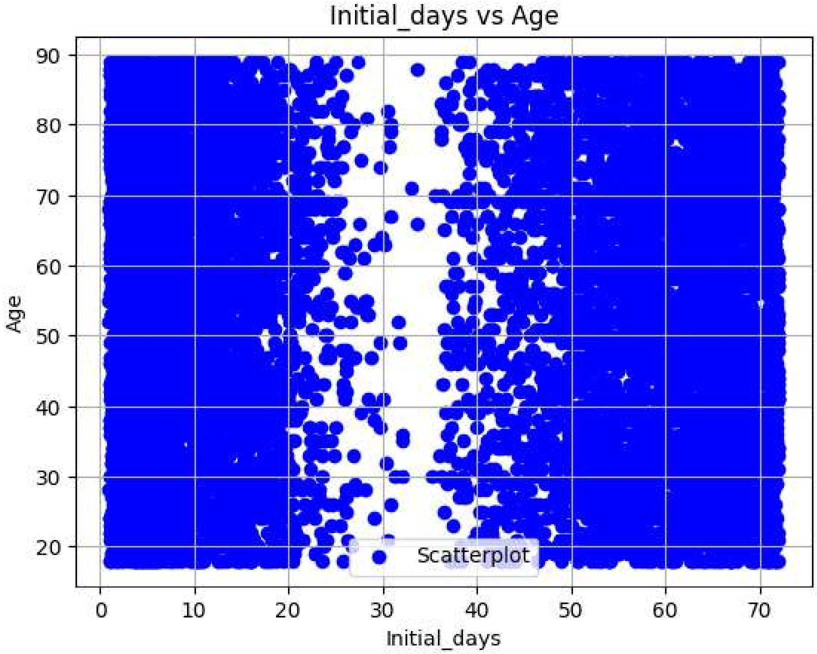

Histogram for VitD_levels
Frequency
0
200
400
600
800
1000
1200
1400
1600
10.0
12.5
15.0
17.5
20.0
22.5
25.0
VitD_levels

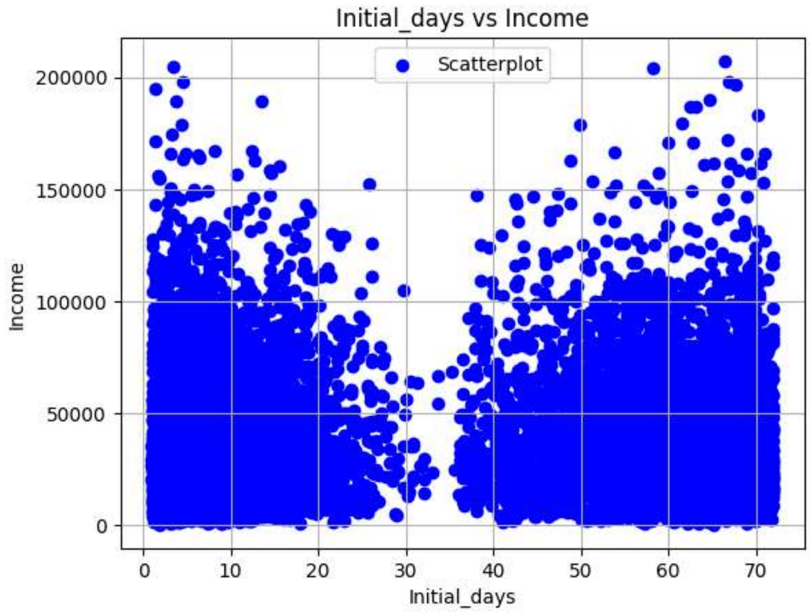

Histogram for Doc_visits
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
1
2
3
4
5
6
7
8
9
Doc_visits

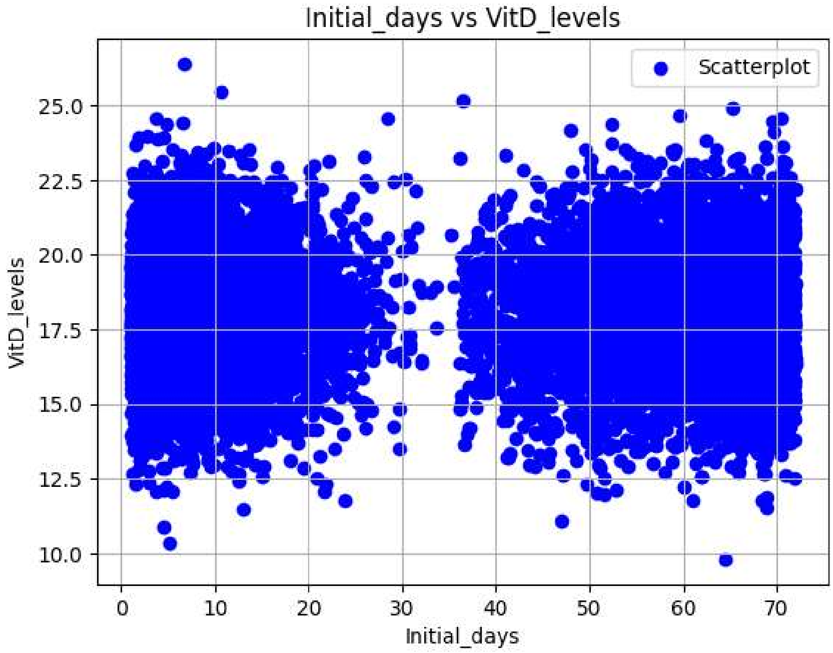

Histogram for vitD_supp
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
vitD_supp

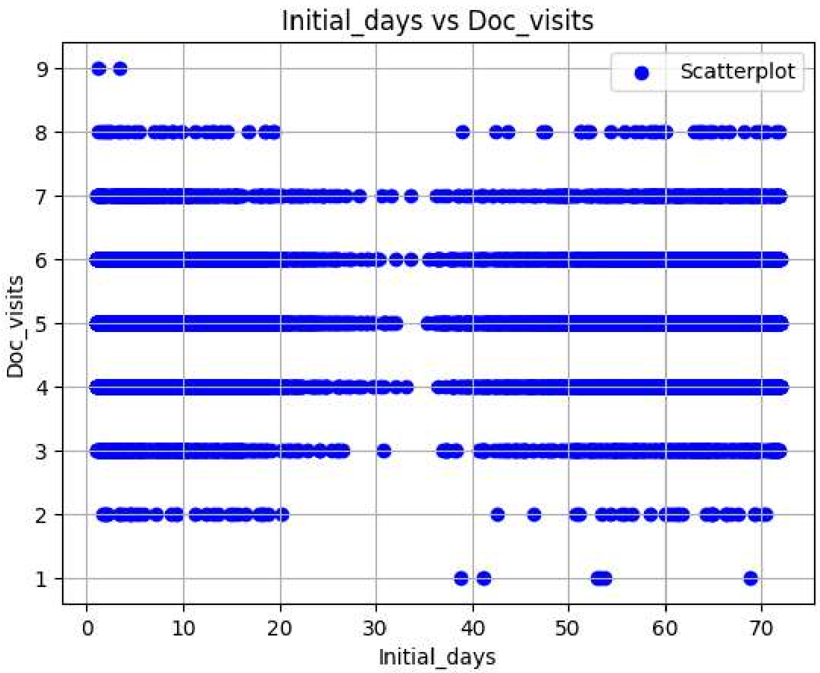
Bivariate visualizations

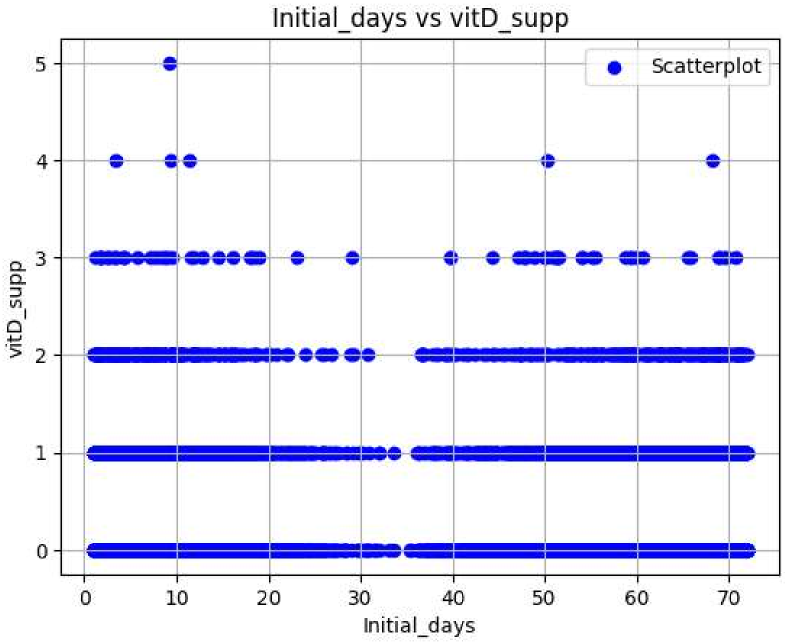





Initial_days vs Income
Scatterplot
Income
200000
150000
100000
50000
0
0
10
20
30
40
50
60
70
Initial_days


Initial_days vs VitD_levels
Scatterplot
VitD_levels
25.0
22.5
20.0
17.5
15.0
12.5
10.0
0
10
20
30
40
50
60
70
Initial_days





## Part C4

The dataset does not have duplicate or missing values, and the outliers are being retained. The dataset will not undergo transformations.

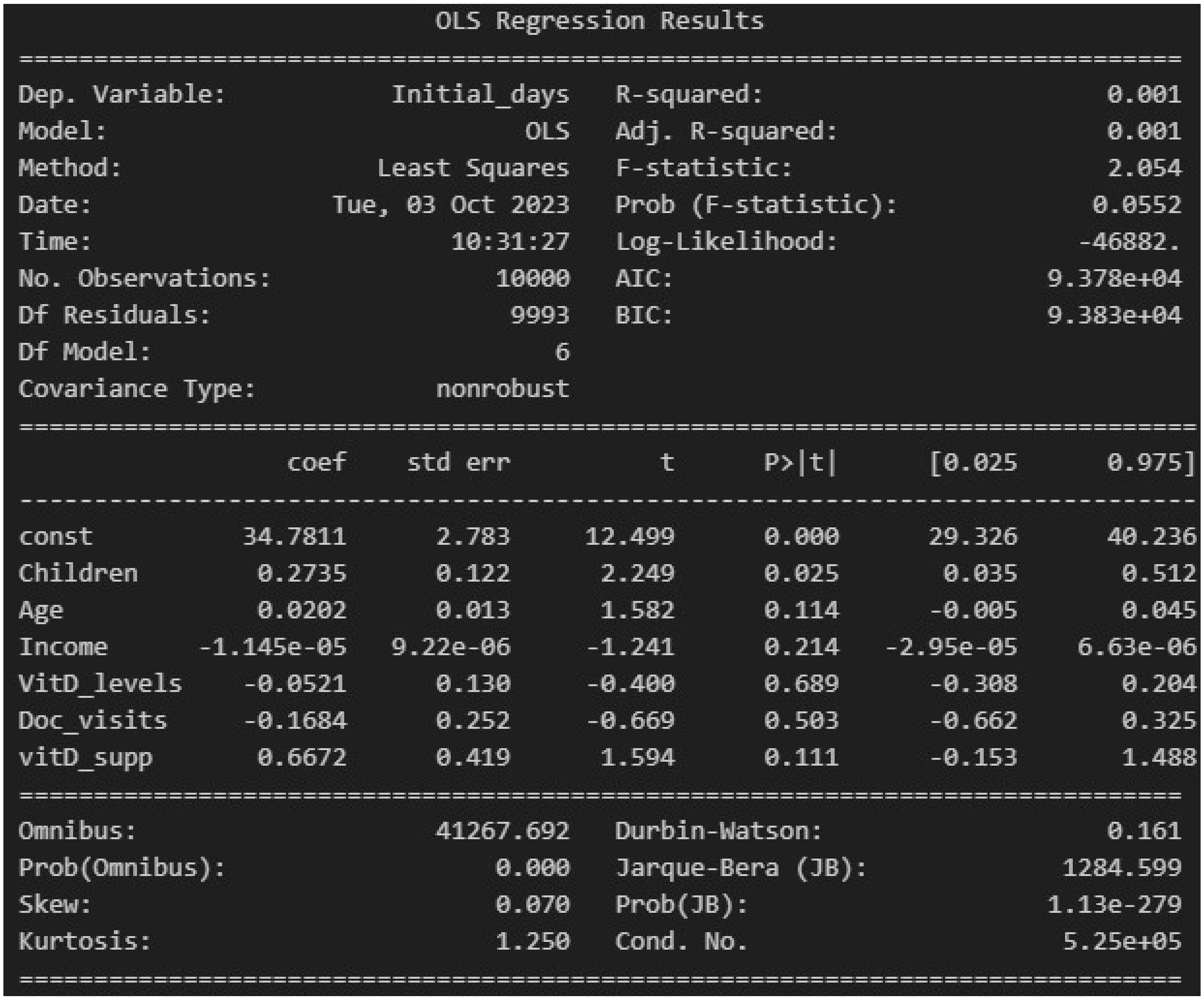
## Part C5

See attached code.

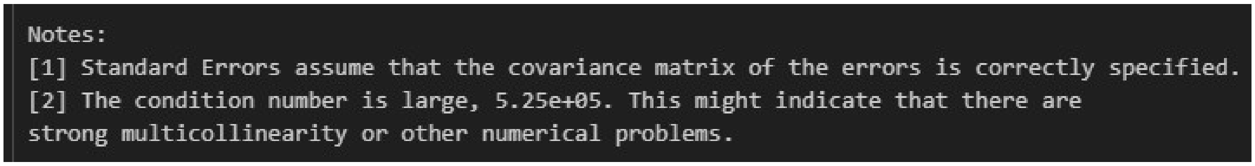
## Part D1

Initial_days = 34.7811 + 0.2735(Children) + 0.0202(Age) – $1.145e^5$(Income) – 0.0521(VitD_levels) – 0.1684(Doc_visits) + 0.6672(vitD_supp)

(GeeksforGeeks, 2023) (Larose & Larose, 2019, sec. 11.4)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     2.054
Date:                Tue, 03 Oct 2023   Prob (F-statistic):             0.0552
Time:                        10:31:27   Log-Likelihood:                -46882.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9993   BIC:                         9.383e+04
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          34.7811      2.783     12.499      0.000      29.326      40.236
Children        0.2735      0.122      2.249      0.025       0.035       0.512
Age             0.0202      0.013      1.582      0.114      -0.005       0.045
Income      -1.145e-05   9.22e-06     -1.241      0.214   -2.95e-05    6.63e-06
VitD_levels    -0.0521      0.130     -0.400      0.689      -0.308       0.204
Doc_visits     -0.1684      0.252     -0.669      0.503      -0.662       0.325
vitD_supp       0.6672      0.419      1.594      0.111      -0.153       1.488
==============================================================================
Omnibus:                    41267.692   Durbin-Watson:                   0.161
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1284.599
Skew:                           0.070   Prob(JB):                    1.13e-279
Kurtosis:                       1.250   Cond. No.                     5.25e+05
==============================================================================
```



```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.25e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

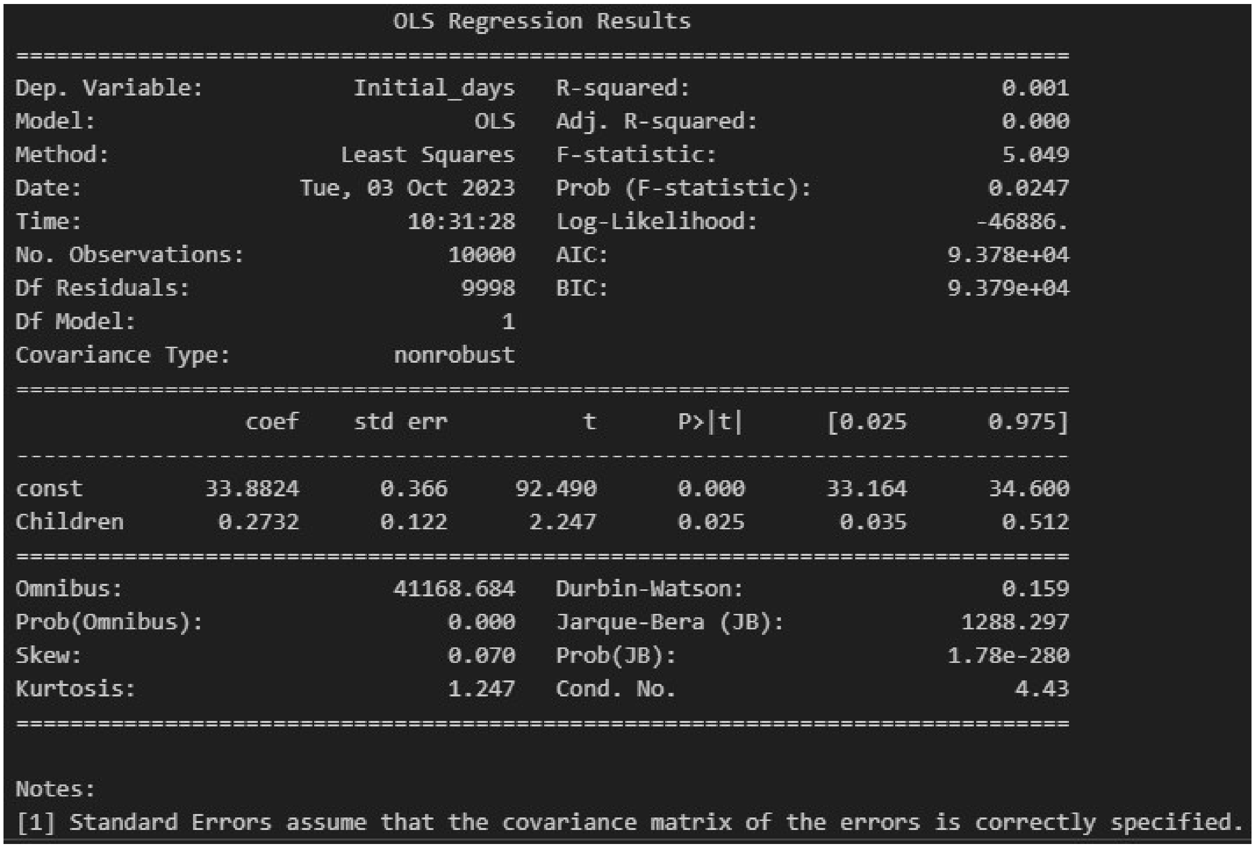
## Part D2

Backward Stepwise Elimination was used as a feature selection procedure to reduce the initial model. This procedure allowed for first evaluating all the possible explanatory variables, and then improving the performance of the model by removing least significant features based on their p-value. This allowed for the model to be evaluated at multiple steps until an acceptable model is achieved.

## Part D3

Initial_days = 33.8824 + 0.2732(Children)

(GeeksforGeeks, 2023) (Larose & Larose, 2019, sec. 11.4)



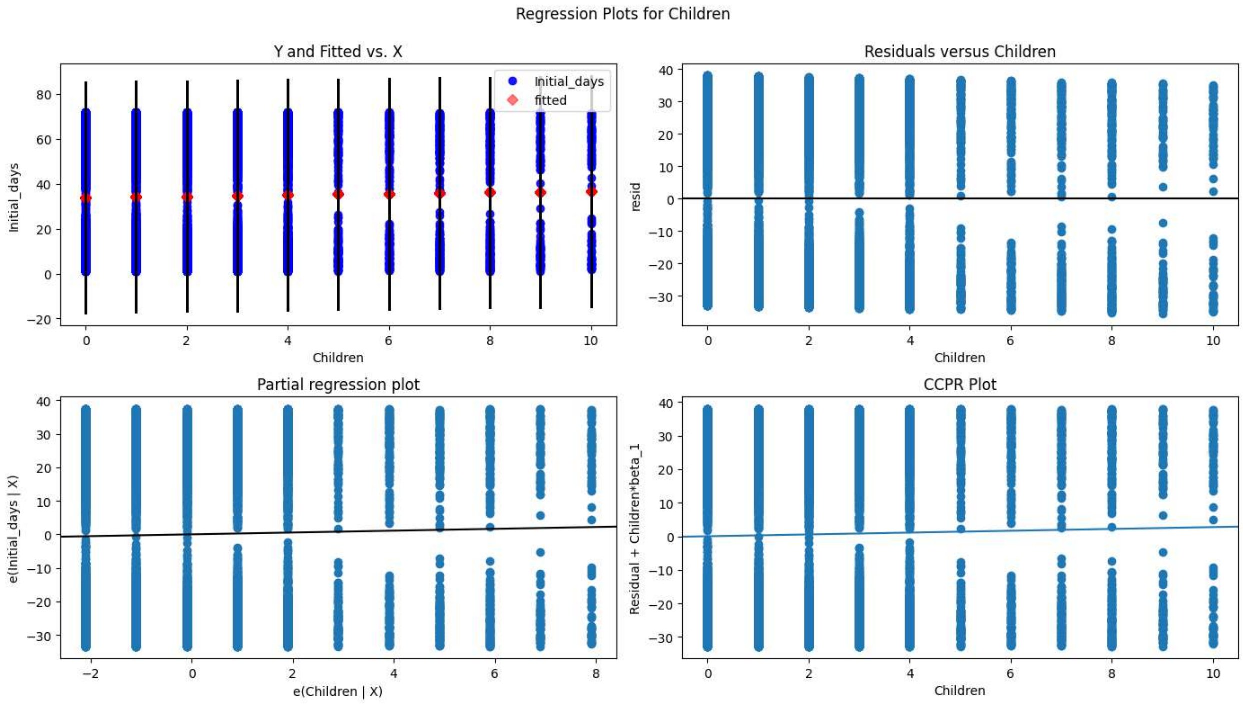
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     5.049
Date:                Tue, 03 Oct 2023   Prob (F-statistic):             0.0247
Time:                        10:31:28   Log-Likelihood:                -46886.
No. Observations:               10000   AIC:                         9.378e+04
Df Residuals:                    9998   BIC:                         9.379e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.8824      0.366     92.490      0.000      33.164      34.600
Children       0.2732      0.122      2.247      0.025       0.035       0.512
==============================================================================
Omnibus:                    41168.684   Durbin-Watson:                   0.159
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1288.297
Skew:                           0.070   Prob(JB):                    1.78e-280
Kurtosis:                       1.247   Cond. No.                         4.43
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

## Part E1

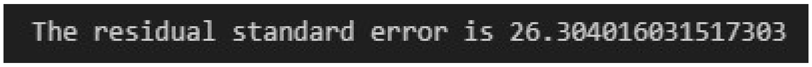
The initial and reduced regression models were evaluated using the Prob(F-statistic) value. The Prob(F-statistic) for the initial model is 0.0552, while the Prob(F-statistic) for the reduced model is 0.0247. The Prob(F-statistic) for the reduced model is less than the p-value, implying that the reduced model is a better fit for the data.

## Part E2

The residual plot was created using the sm.graphics.plot_regress_exog function. (GeeksforGeeks, 2022)



The model’s residual standard error was calculated using the np.sqrt() function. (DSC Data Science Concepts, 2021)



```
The residual standard error is 26.304016031517303
```

## Part E3

See attached code.

## Part F1

The regression equation for the reduced model is Initial_days = 33.8824 + 0.2732(Children).

The coefficient of the reduced model means that as the number of children in the patient’s household increases, the mean of the number of days the patient stayed in the hospital during the initial visit also increases. For every additional child in the patient’s household, the number of days the patient stays in the hospital during the initial visit increases by 0.2732, assuming other factors remain constant.

Statistically, the reduced model has little significance since it eliminated variables, but still did not provide an accurate model. Practically, the reduced model is not significant since it cannot produce a reliable result. The data analysis is limited by the initial selection of explanatory variables. After feature selection was performed, the reduced model cannot accurately fit the data.

## Part F2

Based on my results, I recommend that more explanatory variables be selected before conducting another analysis. This will provide more opportunities for the model to be as accurate as possible.

## Part G

The demonstration can be viewed at https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=815dfc21-af74-4b49-b5e8-b092003f0b16

## Part H

Verma, J. (2020, October 7). How to calculate summary statistics in python?. AskPython. https://www.askpython.com/python/examples/calculate-summary-statistics

GeeksforGeeks. (2023). ML   Multiple Linear Regression using Python. GeeksforGeeks. https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. https://doi.org/10.1002/9781119526865

Stepwise Regression in Python: A Comprehensive guide | Saturn Cloud Blog. (2023, September 9). https://saturncloud.io/blog/stepwise-regression-in-python-a-comprehensive-guide/

GeeksforGeeks. (2022). How to create a residual plot in Python. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-a-residual-plot-in-python/

DSC Data Science Concepts. (2021, November 10). Linear regression. Residual standard error in Python (JuPyter) [Video]. YouTube. https://www.youtube.com/watch?v=QxYmj-E3Ud4

## Part I

None used.