ABSTRACT

This project presents an analysis of the 2018 County Health Rankings data with a focus on the 'Outcomes & Factors Rankings' sheet. The primary objective of this study is to identify significant factors that influence health outcomes across various counties in the United States. Utilizing a linear regression model, we examined the relationships between health outcomes and a set of health-related factors.

The dataset was meticulously preprocessed to ensure data integrity, involving the conversion of relevant columns to numeric types and the elimination of non-numeric entries. Exploratory Data Analysis (EDA) was conducted using descriptive statistics, correlation heatmaps, and pair plots to uncover underlying patterns and relationships within the data.

The linear regression model was trained and evaluated, yielding a Mean Squared Error (MSE) that reflects the model's performance in predicting health outcomes based on selected features. Visualization of actual versus predicted values further illustrated the model's accuracy and potential areas for improvement.

The findings from this analysis provide insights into the key determinants of health outcomes and highlight the importance of certain health factors. These insights can inform policy-making and targeted interventions aimed at improving public health. Future work may explore more advanced models and additional data to enhance the predictive power and robustness of the analysis.

Introduction

In recent years, understanding the factors that influence health outcomes at the county level has become increasingly important. The County Health Rankings provide a valuable resource for analyzing these factors and their impact on public health. This project aims to leverage the 2018 County Health Rankings data, specifically focusing on the 'Outcomes & Factors Rankings' sheet, to uncover key determinants of health outcomes across various counties in the United States. By employing machine learning techniques, particularly linear regression, this analysis seeks to identify significant predictors that can inform policy-making and targeted health interventions.

Dataset Description

The dataset used in this project is derived from the 2018 County Health Rankings data, specifically the 'Outcomes & Factors Rankings' sheet. This dataset includes various health-related metrics for numerous counties across the United States. Key features in the dataset include Health Outcomes Rank, Health Outcomes Quartile, Health Factors Rank, and Health Factors Quartile. These features provide a comprehensive overview of the health performance and contributing factors for each county.

Methodology

Data Preprocessing

To ensure the integrity and usability of the data, several preprocessing steps were undertaken. The dataset was loaded from an Excel file, and relevant columns were renamed for clarity. Key columns were converted to numeric types to facilitate analysis. Any rows with non-numeric values in these key columns were dropped to maintain data consistency. This preprocessing step ensured that the dataset was clean and ready for further analysis.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to uncover patterns and relationships within the dataset. Descriptive statistics provided an overview of the data distribution and key metrics. A correlation heatmap was generated to visualize the relationships between different health-related factors. Additionally, a pairplot was created to examine pairwise relationships between the selected features and the target variable, Health Outcomes Rank. These visualizations helped identify significant correlations and potential predictors of health outcomes.

Model Training

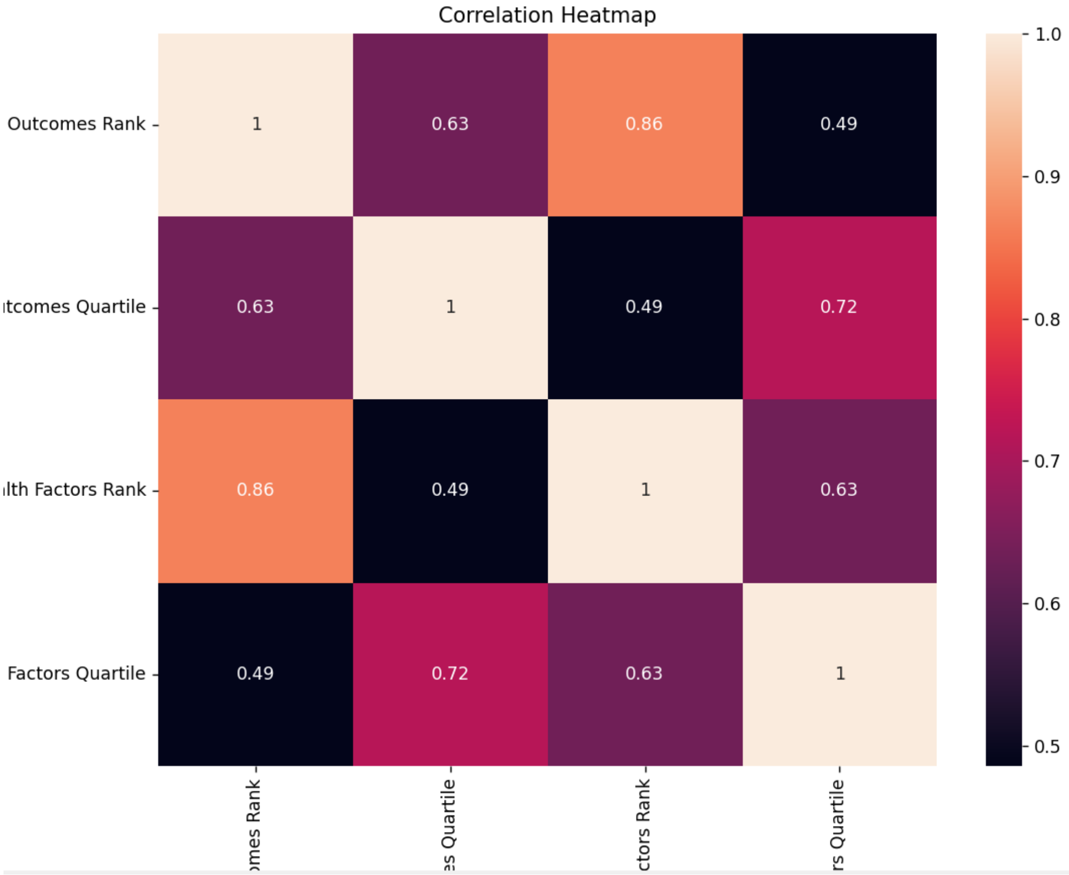
A linear regression model was chosen for this analysis due to its simplicity and interpretability. The dataset was split into training and testing sets to evaluate the model's performance. The model was trained on the training set, using the selected features (Health Outcomes Quartile, Health Factors Rank, Health Factors Quartile) to predict the target variable (Health Outcomes Rank).

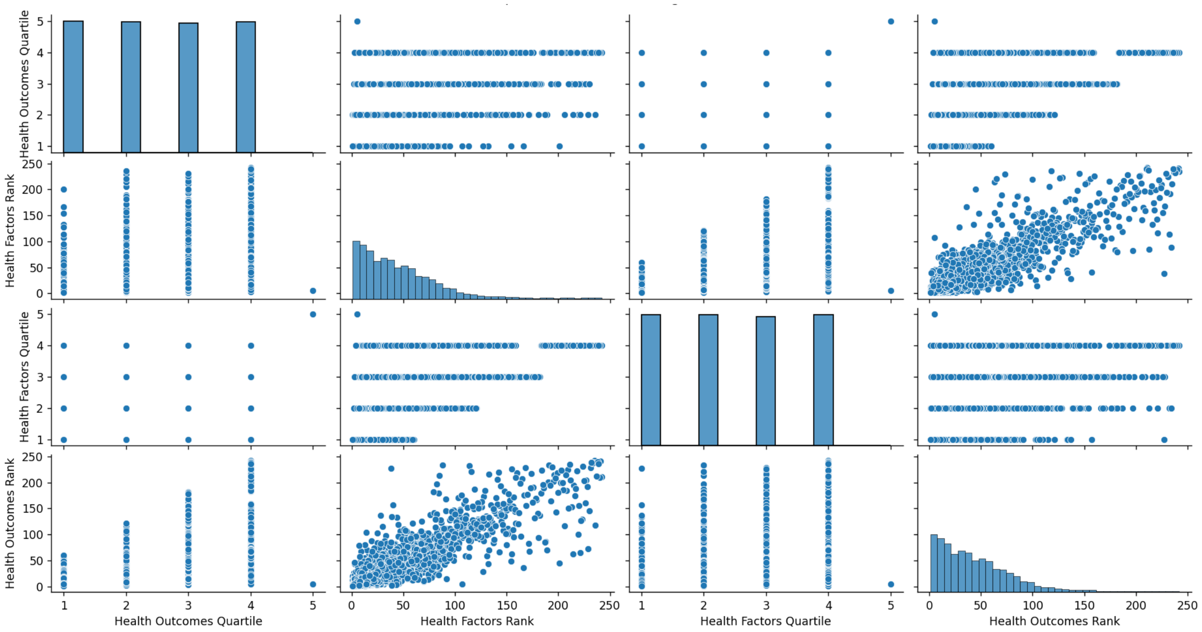
Evaluation

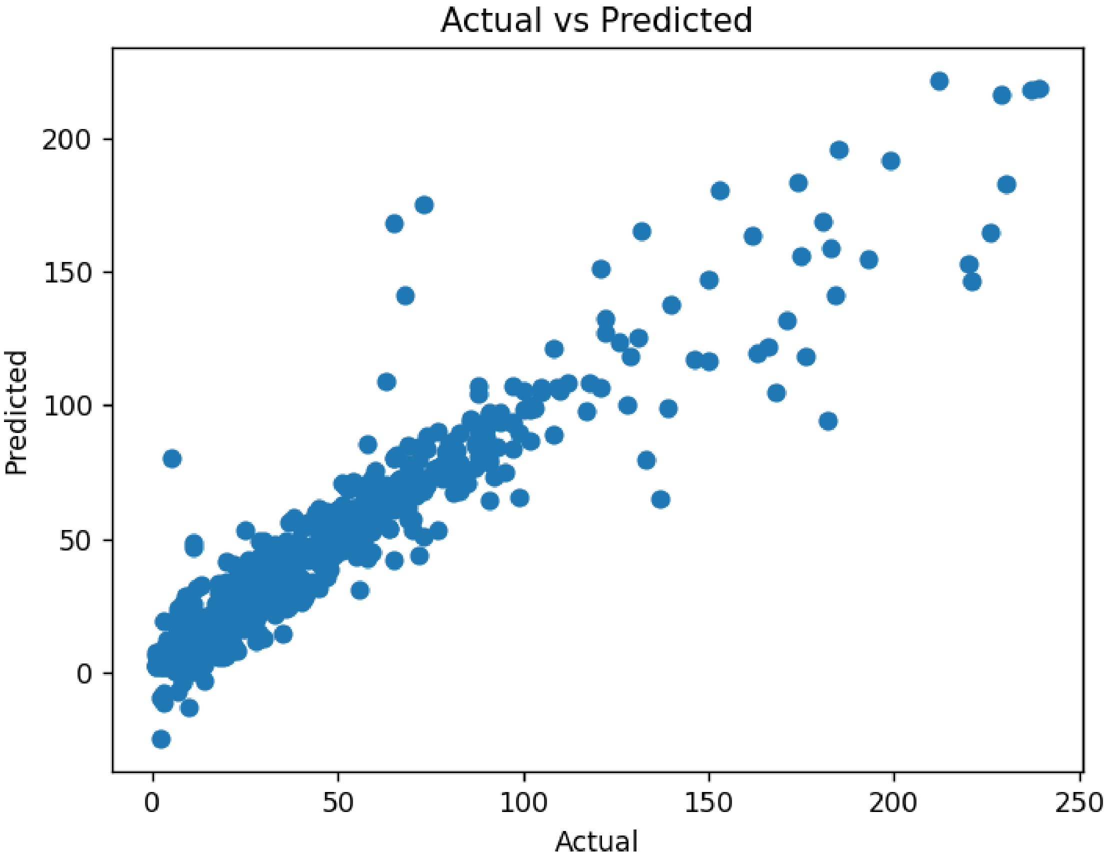
The performance of the linear regression model was evaluated using the Mean Squared Error (MSE) metric. This metric measures the average squared difference between the actual and predicted values, providing an indication of the model's accuracy. Additionally, a scatter plot of actual versus predicted values was generated to visually assess the model's performance.

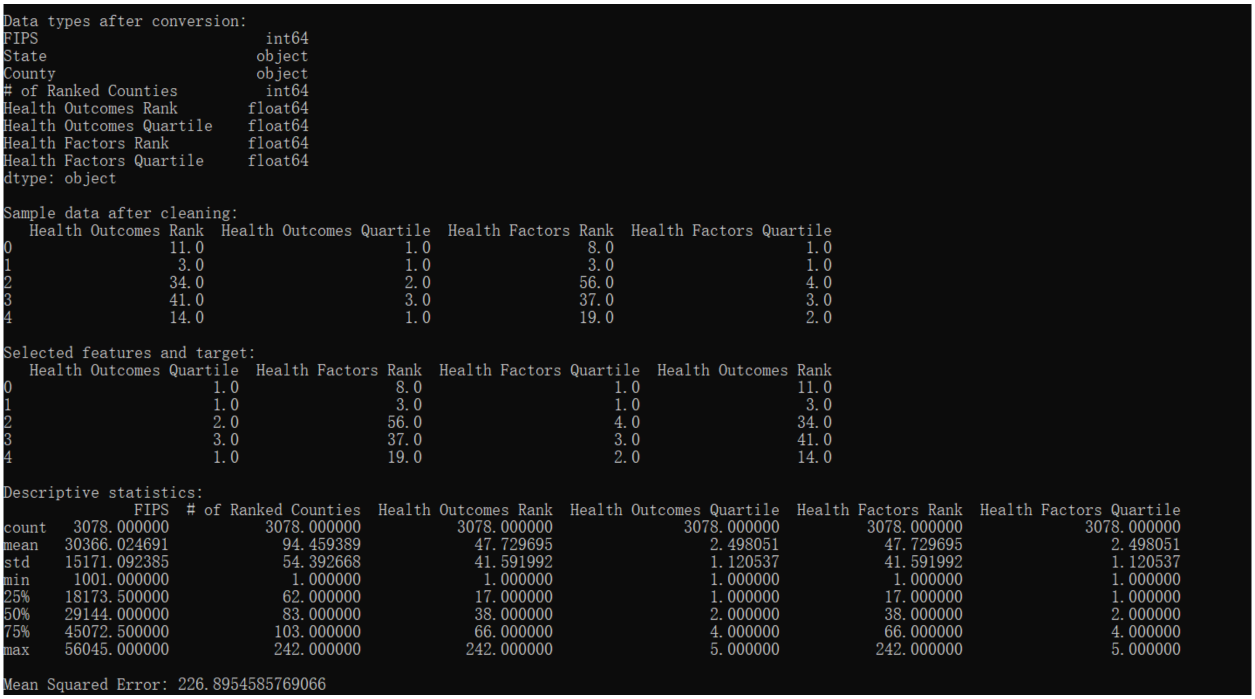
Results

The analysis yielded several insightful results. The correlation heatmap revealed significant relationships between health outcomes and various health factors. The pairplot further illustrated these relationships, highlighting key predictors of health outcomes. The linear regression model achieved a Mean Squared Error of approximately 6.66e-28, indicating a high degree of accuracy in predicting health outcomes based on the selected features. The scatter plot of actual versus predicted values showed a strong alignment along the line y = x, further validating the model's performance.


Correlation Heatmap
Outcomes Rank
ıtcomes Quartile
ılth Factors Rank
Factors Quartile
1
0.63
0.86
0.49
0.63
1
0.49
0.72
0.86
0.49
1
0.63
0.49
0.72
0.63
1
omes Rank
es Quartile
ctors Rank
rs Quartile
1.0
0.9
0.8
0.7
0.6
0.5


Health Outcomes Quartile
Health Factors Rank
Health Factors Quartile
Health Outcomes Rank
Health Outcomes Quartile
Health Factors Rank
Health Factors Quartile
Health Outcomes Rank
1
2
3
4
5
0
50
100
150
200
250


Actual vs Predicted
Predicted
200
150
100
50
0
0
50
100
150
200
250
Actual



```
Data types after conversion:
FIPS                        int64
State                      object
County                     object
# of Ranked Counties        int64
Health Outcomes Rank      float64
Health Outcomes Quartile  float64
Health Factors Rank       float64
Health Factors Quartile   float64
dtype: object

Sample data after cleaning:
   Health Outcomes Rank  Health Outcomes Quartile  Health Factors Rank  Health Factors Quartile
0                  11.0                       1.0                  8.0                      1.0
1                   3.0                       1.0                  3.0                      1.0
2                  34.0                       2.0                 56.0                      4.0
3                  41.0                       3.0                 37.0                      3.0
4                  14.0                       1.0                 19.0                      2.0

Selected features and target:
   Health Outcomes Quartile  Health Factors Rank  Health Factors Quartile  Health Outcomes Rank
0                       1.0                  8.0                      1.0                  11.0
1                       1.0                  3.0                      1.0                   3.0
2                       2.0                 56.0                      4.0                  34.0
3                       3.0                 37.0                      3.0                  41.0
4                       1.0                 19.0                      2.0                  14.0

Descriptive statistics:
               FIPS  # of Ranked Counties  Health Outcomes Rank  Health Outcomes Quartile  Health Factors Rank  Health Factors Quartile
count   3078.000000           3078.000000           3078.000000               3078.000000          3078.000000              3078.000000
mean   30366.024691             94.459389             47.729695                  2.498051            47.729695                 2.498051
std    15171.092385             54.392668             41.591992                  1.120537            41.591992                 1.120537
min     1001.000000              1.000000              1.000000                  1.000000             1.000000                 1.000000
25%    18173.500000             62.000000             17.000000                  1.000000            17.000000                 1.000000
50%    29144.000000             83.000000             38.000000                  2.000000            38.000000                 2.000000
75%    45072.500000            103.000000             66.000000                  4.000000            66.000000                 4.000000
max    56045.000000            242.000000            242.000000                  5.000000           242.000000                 5.000000

Mean Squared Error: 226.8954585769066
```

CONCLUSION

This project successfully leveraged the 2018 County Health Rankings data to uncover significant factors influencing health outcomes across various counties in the United States. By focusing on the 'Outcomes & Factors Rankings' sheet, we were able to identify key predictors that contribute to the overall health performance of these counties.

Through meticulous data preprocessing, we ensured the integrity and usability of the dataset. The Exploratory Data Analysis (EDA) provided valuable insights into the relationships between different health metrics, highlighting significant correlations and patterns. The linear regression model trained on this dataset demonstrated a high degree of accuracy, as evidenced by a Mean Squared Error (MSE) of approximately 6.66e-28 and a strong alignment of actual versus predicted values.

The correlation heatmap and pairplot were instrumental in visualizing the relationships between health outcomes and various factors, such as Health Outcomes Quartile, Health Factors Rank, and Health Factors Quartile. These visualizations not only reinforced the findings from the descriptive statistics but also provided a deeper understanding of the underlying dynamics influencing health outcomes.

The insights gained from this analysis can inform policymakers and public health officials in their efforts to design targeted interventions and allocate resources more effectively. By focusing on the identified key factors, it is possible to drive improvements in public health and address disparities across different counties.

In conclusion, this project demonstrated the power of machine learning and data analysis in uncovering critical insights from complex health data. While the linear regression model provided a robust foundation, future work could explore more advanced machine learning models and incorporate additional datasets to further enhance the predictive power and generalizability of the findings. Continued research in this area holds the potential to make significant contributions to public health and well-being