Jupyter-notebook points:

**Z-Score Method for Outlier Detection**

**Normal Distribution and Z-scores**

In a normal distribution:

* Approximately 68% of the data points lie within 1 standard deviation (±1σ) from the mean.
* Approximately 95% lie within 2 standard deviations (±2σ) from the mean.
* Approximately 99.7% lie within 3 standard deviations (±3σ) from the mean.

**Choosing Thresholds**

* **Z-score Threshold of 3**: This threshold captures 99.7% of the data points if the data follows a normal distribution. It means that only 0.3% of the data points are considered outliers, as they lie beyond ±3σ from the mean.
* **Z-score Threshold of 3.5**: This is a slightly more stringent threshold and captures even more of the central data points. It allows for identifying more extreme outliers, ensuring that only the most extreme 0.023% (approx.) are flagged.

**Why Use These Thresholds?**

1. **Balance**: A threshold of 3 or 3.5 balances between identifying true outliers and not flagging too many data points as outliers, which could distort the analysis.
2. **Simplicity**: These thresholds are simple and widely accepted in statistical practice.
3. **Practicality**: They work well in many real-world datasets, providing a good balance between sensitivity (detecting outliers) and specificity (not flagging normal points).

**Decision Making**

* **Retain Data**: If the differences in distributions and descriptive statistics are minimal, and you prioritize retaining more data, choose the threshold of 3.5.
* **Accuracy**: If you observe significant differences that suggest the presence of impactful outliers in the threshold 3.5 dataset, choose the threshold of 3 for more conservative outlier removal.

Considering the slight variation in weight and the importance of weight as a predictor, you should choose based on the criticality of weight in your model. If weight is a significant factor, opt for the more conservative threshold (3). If you value data retention and the variation is not substantial, the higher threshold (3.5) could be suitable.

**Outlier Detection and Removal**

Outliers can skew the results of your correlation analysis and hypothesis tests. It's important to detect and handle them before scaling.

The sequence in which you perform data preprocessing steps such as outlier detection, hypothesis testing, and scaling can significantly affect the results. Here's a recommended order:

1. **Outlier Detection and Removal**: Detect and potentially remove outliers to ensure that they do not skew your analysis, including scaling and hypothesis testing.
2. **Hypothesis Testing**: Conduct hypothesis tests to analyze the relationships and differences between variables.
3. **Scaling**: Perform scaling after outlier detection and hypothesis testing to standardize your data for further analysis, including correlation analysis and model training.
4. **Hypothesis Testing**
5. After outlier removal, conduct hypothesis testing to analyze the relationships and differences between variables.

#### Scaling

After handling outliers and conducting hypothesis tests, scale your data for further analysis.

**Regression Analysis**

You can perform a linear regression to see if age significantly predicts premium price.

**Formulate Hypotheses**

* H0H\_0H0​: Age is not a significant predictor of premium price.
* H1H\_1H1​: Age is a significant predictor of premium price.

Linear regression and multiple regression are actually part of the same family of models. Linear regression typically refers to simple linear regression, which models the relationship between a dependent variable and a single independent variable. Multiple regression, on the other hand, models the relationship between a dependent variable and multiple independent variables.

Given that you have multiple potential predictors (features) for the insurance premium price, multiple regression would be the appropriate choice. Here's why:

1. **Capturing Complexity**: Multiple regression can capture the combined effect of multiple predictors on the dependent variable, providing a more comprehensive understanding of the data.
2. **Feature Importance**: It allows you to understand the relative importance of each predictor.
3. **Model Performance**: Multiple regression often results in better model performance (e.g., higher R-squared value) compared to simple linear regression when there are multiple predictors involved.

# Problem Statement: Insurance companies need to accurately predict the cost of health insurance for individuals to

# set premiums appropriately. However, traditional methods of cost prediction often rely on broad actuarial tables

# and historical averages, which may not account for the nuanced differences among individuals. By leveraging

# machine learning techniques, insurers can predict more accurately the insurance costs tailored to individual profiles,

# leading to more competitive pricing and better risk management.

Findings:

# The dataset has 986 rows and 11 columns. The columns have no null values in them.

#The above table shows statistical analysis data for each column having its mean value, minimum value,

# its 25,50 and 75th percentile values and then maximum values. Also count of how many data points in each column.

# There is high correlation between 1) Age-Premium Price and 2) Age - Number of major surgeries as shown in above heatmap.

Number of outliers with threshold 3: 29

Number of outliers with threshold 3.5: 5

# Outlier threshold 3 is chosen to give more accurate predictions on premiums from the other variables given.

#Above histogram plots are comparison plots of same variable with different threshold of 3 and 3.5 and are found to be almost similar.

#There is high correlation between Age-Premium Price only now after scaling the data.

Hypothesis testing:

1. Pearson Correlation Coefficient: 0.7015949949405942

P-value: 1.0447703338239117e-142

Reject the null hypothesis: There is a significant correlation between age and premium price.

in scatter plot it shows roughly as age increases the premiums also increase but is same for certain age groups.

1. T-statistic: 2.1083874439817247

P-value: 0.03525745245686358

Reject the null hypothesis: There is a significant difference in premium prices between individuals with and without diabetes.

1. T-statistic: 5.349802905099321

P-value: 1.1024518792137634e-07

Reject the null hypothesis: There is a significant difference in premium prices between individuals with and without BP Problems.

1. T-statistic: 9.329781309507613

P-value: 7.19294884464435e-20

Reject the null hypothesis: There is a significant difference in premium prices between individuals with and without Transplants.

1. T-statistic: 6.904386321211786

P-value: 9.186718789391516e-12

Reject the null hypothesis: There is a significant difference in premium prices between individuals with and without Chronic Diseases.

1. T-statistic: 0.33484693206858634

P-value: 0.7378140856998932

Fail to reject the null hypothesis: There is no significant difference in premium prices between individuals with and without Known Allergies.

1. T-statistic: 2.5914669114379043 (ttest-ind)

P-value: 0.009702608654041739

Reject the null hypothesis: There is a significant difference in premium prices between individuals with and without History of cancer in family.

1. F-statistic: 26.13539359740762(f-oneway)

P-value: 2.8711631377228097e-16

Reject the null hypothesis: There is a significant difference in premium prices across different numbers of surgeries.

# It is found that there is no significant difference in premiums for known allergies whether 0 or 1 but other health factors

# do have significant difference in premiums especially for those whom there is presence of that particular health condition.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

==========================================================

group1 group2 meandiff p-adj lower upper reject

----------------------------------------------------------

0 1 1773.2507 0.0001 702.2872 2844.2142 True

0 2 5115.3489 0.0 3528.0508 6702.6469 True

0 3 5031.3152 0.0058 1092.8776 8969.7529 True

1 2 3342.0981 0.0 1710.0051 4974.1912 True

1 3 3258.0645 0.1477 -698.6392 7214.7682 False

2 3 -84.0336 0.9999 -4210.5402 4042.4729 False

----------------------------------------------------------

Only in case of Known allergies there is no change in premium price whether it is there or not. But for other health factors there are difference in premium price when it is there versus than it is not there. And obviously the Premium price will be more for when these health factors are present in a person.

# Mean of premium = 28084.03361344538 and standard of premium = 1175.8864808346564 after grouping of data of major surgeries to further find outliers in them.

Contingency Table for Diabetes and BloodPressureProblems:

BloodPressureProblems 0 1

Diabetes

0 335 237

1 189 225

Chi-square Test Results for Diabetes and BloodPressureProblems:

Chi-square statistic: 15.571225392078397

P-value: 7.945461025980538e-05

Degrees of freedom: 1

Expected frequencies:

[[303.98377282 268.01622718]

[220.01622718 193.98377282]]

Reject the null hypothesis: There is a significant association between Diabetes and BloodPressureProblems.

Contingency Table for Diabetes and AnyTransplants:

AnyTransplants 0 1

Diabetes

0 536 36

1 395 19

Chi-square Test Results for Diabetes and AnyTransplants:

Chi-square statistic: 1.0207290599382488

P-value: 0.3123461364949182

Degrees of freedom: 1

Expected frequencies:

[[540.09330629 31.90669371]

[390.90669371 23.09330629]]

Fail to reject the null hypothesis: There is no significant association between Diabetes and AnyTransplants.

Contingency Table for Diabetes and AnyChronicDiseases:

AnyChronicDiseases 0 1

Diabetes

0 452 120

1 356 58

Chi-square Test Results for Diabetes and AnyChronicDiseases:

Chi-square statistic: 7.421404393184893

P-value: 0.006445248971042237

Degrees of freedom: 1

Expected frequencies:

[[468.73833671 103.26166329]

[339.26166329 74.73833671]]

Reject the null hypothesis: There is a significant association between Diabetes and AnyChronicDiseases.

Contingency Table for Diabetes and KnownAllergies:

KnownAllergies 0 1

Diabetes

0 433 139

1 341 73

Chi-square Test Results for Diabetes and KnownAllergies:

Chi-square statistic: 5.937672731640452

P-value: 0.014820575766547196

Degrees of freedom: 1

Expected frequencies:

[[449.01419878 122.98580122]

[324.98580122 89.01419878]]

Reject the null hypothesis: There is a significant association between Diabetes and KnownAllergies.

Contingency Table for Diabetes and HistoryOfCancerInFamily:

HistoryOfCancerInFamily 0 1

Diabetes

0 496 76

1 374 40

Chi-square Test Results for Diabetes and HistoryOfCancerInFamily:

Chi-square statistic: 2.7008950288841564

P-value: 0.10029193010869264

Degrees of freedom: 1

Expected frequencies:

[[504.70588235 67.29411765]

[365.29411765 48.70588235]]

Fail to reject the null hypothesis: There is no significant association between Diabetes and HistoryOfCancerInFamily.

Contingency Table for Diabetes and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

Diabetes

0 282 244 45 1

1 197 128 74 15

Chi-square Test Results for Diabetes and NumberOfMajorSurgeries:

Chi-square statistic: 46.44698206025138

P-value: 4.5566988133245854e-10

Degrees of freedom: 3

Expected frequencies:

[[277.87829615 215.80527383 69.03448276 9.28194726]

[201.12170385 156.19472617 49.96551724 6.71805274]]

Reject the null hypothesis: There is a significant association between Diabetes and NumberOfMajorSurgeries.

Contingency Table for BloodPressureProblems and AnyTransplants:

AnyTransplants 0 1

BloodPressureProblems

0 492 32

1 439 23

Chi-square Test Results for BloodPressureProblems and AnyTransplants:

Chi-square statistic: 0.39874887344149174

P-value: 0.5277360989921708

Degrees of freedom: 1

Expected frequencies:

[[494.77079108 29.22920892]

[436.22920892 25.77079108]]

Fail to reject the null hypothesis: There is no significant association between BloodPressureProblems and AnyTransplants.

Contingency Table for BloodPressureProblems and AnyChronicDiseases:

AnyChronicDiseases 0 1

BloodPressureProblems

0 438 86

1 370 92

Chi-square Test Results for BloodPressureProblems and AnyChronicDiseases:

Chi-square statistic: 1.8046980449571972

P-value: 0.17914556026705172

Degrees of freedom: 1

Expected frequencies:

[[429.40365112 94.59634888]

[378.59634888 83.40365112]]

Fail to reject the null hypothesis: There is no significant association between BloodPressureProblems and AnyChronicDiseases.

Contingency Table for BloodPressureProblems and KnownAllergies:

KnownAllergies 0 1

BloodPressureProblems

0 409 115

1 365 97

Chi-square Test Results for BloodPressureProblems and KnownAllergies:

Chi-square statistic: 0.08122766415511057

P-value: 0.7756405611001655

Degrees of freedom: 1

Expected frequencies:

[[411.3346856 112.6653144]

[362.6653144 99.3346856]]

Fail to reject the null hypothesis: There is no significant association between BloodPressureProblems and KnownAllergies.

Contingency Table for BloodPressureProblems and HistoryOfCancerInFamily:

HistoryOfCancerInFamily 0 1

BloodPressureProblems

0 470 54

1 400 62

Chi-square Test Results for BloodPressureProblems and HistoryOfCancerInFamily:

Chi-square statistic: 2.0041718300783176

P-value: 0.1568669440083543

Degrees of freedom: 1

Expected frequencies:

[[462.35294118 61.64705882]

[407.64705882 54.35294118]]

Fail to reject the null hypothesis: There is no significant association between BloodPressureProblems and HistoryOfCancerInFamily.

Contingency Table for BloodPressureProblems and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

BloodPressureProblems

0 315 172 26 11

1 164 200 93 5

Chi-square Test Results for BloodPressureProblems and NumberOfMajorSurgeries:

Chi-square statistic: 86.12341484802883

P-value: 1.4893488650667128e-18

Degrees of freedom: 3

Expected frequencies:

[[254.55983773 197.69574037 63.24137931 8.5030426 ]

[224.44016227 174.30425963 55.75862069 7.4969574 ]]

Reject the null hypothesis: There is a significant association between BloodPressureProblems and NumberOfMajorSurgeries.

Contingency Table for AnyTransplants and AnyChronicDiseases:

AnyChronicDiseases 0 1

AnyTransplants

0 766 165

1 42 13

Chi-square Test Results for AnyTransplants and AnyChronicDiseases:

Chi-square statistic: 0.8603776272400288

P-value: 0.35363303686928205

Degrees of freedom: 1

Expected frequencies:

[[762.92900609 168.07099391]

[ 45.07099391 9.92900609]]

Fail to reject the null hypothesis: There is no significant association between AnyTransplants and AnyChronicDiseases.

Contingency Table for AnyTransplants and KnownAllergies:

KnownAllergies 0 1

AnyTransplants

0 731 200

1 43 12

Chi-square Test Results for AnyTransplants and KnownAllergies:

Chi-square statistic: 0.0

P-value: 1.0

Degrees of freedom: 1

Expected frequencies:

[[730.82555781 200.17444219]

[ 43.17444219 11.82555781]]

Fail to reject the null hypothesis: There is no significant association between AnyTransplants and KnownAllergies.

Contingency Table for AnyTransplants and HistoryOfCancerInFamily:

HistoryOfCancerInFamily 0 1

AnyTransplants

0 820 111

1 50 5

Chi-square Test Results for AnyTransplants and HistoryOfCancerInFamily:

Chi-square statistic: 0.17474758324382383

P-value: 0.6759264929226855

Degrees of freedom: 1

Expected frequencies:

[[821.47058824 109.52941176]

[ 48.52941176 6.47058824]]

Fail to reject the null hypothesis: There is no significant association between AnyTransplants and HistoryOfCancerInFamily.

Contingency Table for AnyTransplants and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

AnyTransplants

0 453 349 114 15

1 26 23 5 1

Chi-square Test Results for AnyTransplants and NumberOfMajorSurgeries:

Chi-square statistic: 0.7205169893162533

P-value: 0.8683678631155989

Degrees of freedom: 3

Expected frequencies:

[[452.28093306 351.2494929 112.36206897 15.10750507]

[ 26.71906694 20.7505071 6.63793103 0.89249493]]

Fail to reject the null hypothesis: There is no significant association between AnyTransplants and NumberOfMajorSurgeries.

Contingency Table for AnyChronicDiseases and KnownAllergies:

KnownAllergies 0 1

AnyChronicDiseases

0 630 178

1 144 34

Chi-square Test Results for AnyChronicDiseases and KnownAllergies:

Chi-square statistic: 0.5778579425673225

P-value: 0.4471531660063278

Degrees of freedom: 1

Expected frequencies:

[[634.27180527 173.72819473]

[139.72819473 38.27180527]]

Fail to reject the null hypothesis: There is no significant association between AnyChronicDiseases and KnownAllergies.

Contingency Table for AnyChronicDiseases and HistoryOfCancerInFamily:

HistoryOfCancerInFamily 0 1

AnyChronicDiseases

0 714 94

1 156 22

Chi-square Test Results for AnyChronicDiseases and HistoryOfCancerInFamily:

Chi-square statistic: 0.02062393388215223

P-value: 0.8858081638149811

Degrees of freedom: 1

Expected frequencies:

[[712.94117647 95.05882353]

[157.05882353 20.94117647]]

Fail to reject the null hypothesis: There is no significant association between AnyChronicDiseases and HistoryOfCancerInFamily.

Contingency Table for AnyChronicDiseases and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

AnyChronicDiseases

0 396 305 91 16

1 83 67 28 0

Chi-square Test Results for AnyChronicDiseases and NumberOfMajorSurgeries:

Chi-square statistic: 6.108071508903402

P-value: 0.10646907184162867

Degrees of freedom: 3

Expected frequencies:

[[392.52738337 304.84381339 97.51724138 13.11156187]

[ 86.47261663 67.15618661 21.48275862 2.88843813]]

Fail to reject the null hypothesis: There is no significant association between AnyChronicDiseases and NumberOfMajorSurgeries.

Contingency Table for KnownAllergies and HistoryOfCancerInFamily:

HistoryOfCancerInFamily 0 1

KnownAllergies

0 698 76

1 172 40

Chi-square Test Results for KnownAllergies and HistoryOfCancerInFamily:

Chi-square statistic: 12.26956724440544

P-value: 0.0004604050250651458

Degrees of freedom: 1

Expected frequencies:

[[682.94117647 91.05882353]

[187.05882353 24.94117647]]

Reject the null hypothesis: There is a significant association between KnownAllergies and HistoryOfCancerInFamily.

Contingency Table for KnownAllergies and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

KnownAllergies

0 426 227 105 16

1 53 145 14 0

Chi-square Test Results for KnownAllergies and NumberOfMajorSurgeries:

Chi-square statistic: 109.30170166278893

P-value: 1.5509963903476643e-23

Degrees of freedom: 3

Expected frequencies:

[[376.01014199 292.01622718 93.4137931 12.55983773]

[102.98985801 79.98377282 25.5862069 3.44016227]]

Reject the null hypothesis: There is a significant association between KnownAllergies and NumberOfMajorSurgeries.

Contingency Table for HistoryOfCancerInFamily and NumberOfMajorSurgeries:

NumberOfMajorSurgeries 0 1 2 3

HistoryOfCancerInFamily

0 479 268 107 16

1 0 104 12 0

Chi-square Test Results for HistoryOfCancerInFamily and NumberOfMajorSurgeries:

Chi-square statistic: 160.28223246287763

P-value: 1.5930259415728302e-34

Degrees of freedom: 3

Expected frequencies:

[[422.64705882 328.23529412 105. 14.11764706]

[ 56.35294118 43.76470588 14. 1.88235294]]

Reject the null hypothesis: There is a significant association between HistoryOfCancerInFamily and NumberOfMajorSurgeries.

# 8 significant association between health factors were found from the chi tests.

#linear regression proves that age is a significant predictor of premium price.

Significant predictors: ['const', 'Age', 'Weight'] found from multiple regression analysis.

Model Selection: Random Forest:

Mean Squared Error: 4529460.101010101

R-squared: 0.8937813417890521

# Result shows Random Forest is the best with lowest MSE and highest R squared value among other tests done.

Hyperparameter tuning:

Fitting 5 folds for each of 36 candidates, totalling 180 fits

Best parameters for Random Forest: {'max\_depth': None, 'min\_samples\_split': 10, 'n\_estimators': 300}

Mean Squared Error: 4127369.9325902294

R-squared: 0.903210606473346

#Age has the highest importance, indicating it is the most significant predictor for premium price.

#AnyTransplants and Weight also have considerable importance, suggesting their influence on premium prices.

#Features like BMI, AnyChronicDiseases, and NumberOfMajorSurgeries have moderate impact.

#Features like Diabetes and KnownAllergies have minimal importance.

Model Evaluation and validation:

Cross-Validation RMSE Scores: [2226.37154287 1778.09540797 2672.32339681 3764.97471746 2952.97476448

3914.92533882 2955.8645863 3084.10544703 1725.88216319 2718.05194717]

Mean RMSE: 2779.3569312082736

Standard Deviation of RMSE: 696.3094738854265

Model performance metrics:

Mean Squared Error: 3633369.8954909276

Root Mean Squared Error: 1906.1400513841913

Mean Absolute Error: 698.9681068545847

R-squared: 0.9068371804247812

Condidence Interval/Prediction Interval for instance 0: (23494.699938160964, 24973.831082561235)

**Model Interpretation**:

* **Age** has the highest importance, indicating it is the most significant predictor for premium price.
* **AnyTransplants** and **Weight** also have considerable importance, suggesting their influence on premium prices.
* Features like **BMI**, **AnyChronicDiseases**, and **NumberOfMajorSurgeries** have moderate impact.
* Features like **Diabetes** and **KnownAllergies** have minimal importance.

**Interpretation of SHAP Dependence Plots:**

1. **Positive SHAP Values**: Indicate an increase in the premium price.
2. **Negative SHAP Values**: Indicate a decrease in the premium price.
3. **Color Gradient**: Helps to see interaction effects. For example, if the color represents another feature, you can see how two features interact to affect the prediction.
4.  **Cross-Validation**: Helps in assessing model stability and generalizability.
5.  **Performance Metrics**: Evaluate the model's accuracy and error.
6.  **Confidence/Prediction Intervals**: Provide uncertainty estimates for predictions.

Prediction intervals are useful for understanding the reliability of your model’s predictions and can help in decision-making by quantifying the expected range of outcomes. They are particularly valuable in scenarios where knowing the range of potential values is crucial for risk assessment and decision-making.

The Mean RMSE gives an average error across all cross-validation folds. A lower RMSE indicates better model performance. The standard deviation shows how much the RMSE varies across different subsets of the data. A low standard deviation means the model is consistent across different data splits.

Performance metrics:

* **MSE and RMSE**: Lower values indicate better model performance. RMSE is easier to interpret as it’s in the same units as the target variable.
* **MAE**: Provides an average of absolute errors. A lower MAE indicates better model performance.
* **R-squared**: This value indicates how well the model explains the variance in the target variable. An R-squared of 0.91 suggests that the model explains 91% of the variance, which is excellent.

**Evaluation**:

* **RMSE**: Compare this with the range of premium prices. If it's small relative to the values, your model is performing well.
* **MAE**: Provides an easy-to-understand measure of average error.
* **R-squared**: A value close to 1.0 indicates a good fit, suggesting your model is explaining most of the variance.

**Summary**

**Overall**:

* The RMSE, MAE, and R-squared values indicate that the model is performing well.
* The prediction interval provides a useful range for understanding the uncertainty around predictions.

#Insights based on the importance of each feature:

#Age: Older individuals are likely to have higher insurance premiums.

#Insight: Targeted health programs and regular check-ups can be introduced for older clients to manage health risks better.

#BMI: Higher BMI is associated with higher premiums.

#Insight: Offer wellness programs and incentives for maintaining a healthy BMI to reduce future claims.

#Weight: Higher weight contributes to higher premiums.

#Insight: Weight management programs can be offered to customers.

#Number of Major Surgeries: More surgeries correlate with higher premiums.

#Insight: Monitor and support customers with a history of major surgeries with preventive care.

#Any Transplants: Individuals with transplants have higher premiums.

#Insight: Provide specialized care and regular follow-ups for transplant recipients.

#Health Monitoring Programs: Develop programs for regular health check-ups, especially for older customers,

# to identify and manage health risks early.

#Wellness and Fitness Initiatives: Provide incentives for customers to participate in fitness programs and

# maintain a healthy BMI.

#Preventive Care Plans: Offer comprehensive preventive care plans for customers with a history of surgeries and transplants

# to reduce the likelihood of future claims.

#Personalized Insurance Plans: Customize insurance plans based on individual risk factors to provide fair and

# competitive premiums.

#Summary

#By analyzing the feature importances from the model, we can identify key risk factors contributing to

#higher insurance premiums.

#Targeted interventions based on these insights can help manage risks more effectively and provide better value to customers.