**Dandi Mahendris - dandi-P4Gx**

**-----**

Deductible Amount Consideration of Policy Holder Towards Several Health Conditions

CONTENTS

1. **Introduction 1**
   1. Definitions 1
   2. Objectives 1
2. **Datasets 1**
3. **Exploratory Analysis 2**

3.1. Distribution of Inpatient-Outpatient Reimbursement and Deductible Amount 2

3.2. Proportion of all variable 3

3.3. Distribution of Variable toward Outpatient Deductible Amount 3

3.4. Correlation between OP Deductible Amount and Numerical Variable 5

3.5. OP Deductible against OP Reimbursement amount 5

1. **Statistical Inference 6**
2. **Model Fitting 9**
3. **Limitation of Work 10**
4. **Conclusions 10**
5. **Introduction**

Deductibles have been an essential part of the insurance contract for many years. Understanding the role deductibles play when insurer life, vehicle or home is integral to getting the most out of your insurance policy.

Insurance deductible pertains to the amount of money on an insurance claim that you would pay before the coverage kicks in and the insurer pays. In other words, it is the money that you would shell out of your own pocket before receiving insurance coverage. After paying your deductible, the insurance company will start paying the remaining amount of the claim value up to the limits indicated in the policy.

The larger the deductible, the less you pay in premiums for an insurance policy. A deductible can be either a specific dollar amount or a percentage of the total amount of insurance on a policy. The amount is established by the terms of your coverage and can be found on the declarations (or front) page of standard homeowners, condo owners, renters, and auto insurance policies.

**1.1 Definitions**

**1. Deductible:**

Amount of money on an insurance claim that you would pay before receiving insurance coverage or insurer start to pays.

**2. Reimbursement:**

payments made by insurer to a policy holder who has made a claim for expenses incurred out of their own pocket for the medical expense.

**2. Inpatient:**

a patient who receives medical treatment that requires to be admitted to a hospital or other care facility for at least one night.

**3. Outpatient:**

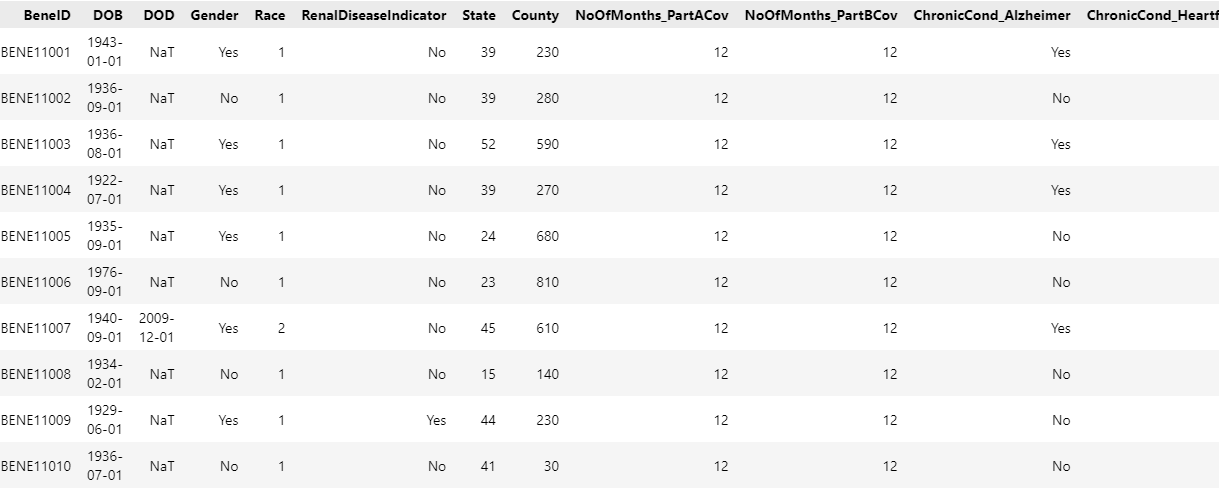
a patient who receives medical treatment without being admitted to a hospital or other care facility.

* 1. **Objective**

1. Which variable is being highly consideration of someone to choose low or high outpatient deductibility?

1. Does Renal Disease Indicator affect the consideration of outpatient deductibility?
2. Does Gender have affected outpatient deductibility?
3. How much does Kidney, Stroke, Alzheimer, Heart failure, Cancer, Diabetes condition affect outpatient deductibility?
4. How is relationship among variables?
5. **Dataset**

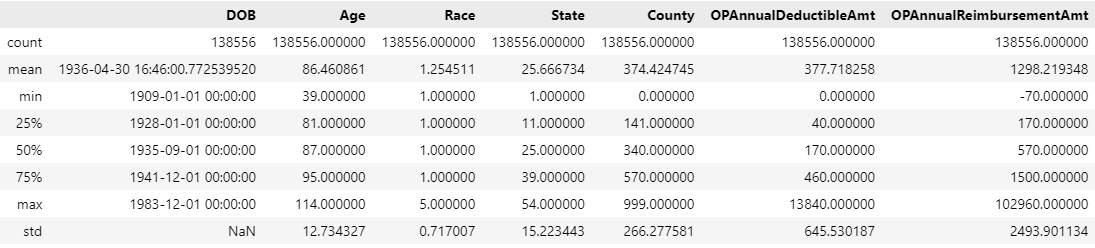
Data was collected in Kaggle, as one of dataset for Healthcare Provider contains beneficiary KYC details like health conditions of few diseases, Date of Birth, Gender, Race, State, Countries, etc. Total data row 138,556 and total 25 columns.



Variable Descriptions:

1. **BeneId**: Beneficiary ID or person who receives policy benefit.
2. **DOB:** Date of Birth
3. **DOD:** Date of Death
4. **Gender:** Gender (Male or Female)
5. **Renal Disease Indicator:** Renal disease indicator to kidney disease potential (1 = No; 2= Yes)
6. **ChronicCond\_Alzheimer:** Have an alzheimer medical record? (1 = No; 2= Yes)
7. **ChronicCond\_Heartfailure:** Have a heart failure medical record? (1 = No; 2= Yes)
8. **ChronicCond\_KidneyDisease:** Have a kidney disease medical record? (1 = No; 2= Yes)
9. **ChronicCond\_Cancer:** Have a cancer medical record? (1 = No; 2= Yes)
10. **ChronicCond\_ObstrPulmonary:** Have an obstr pulmonary medical record? (1 = No; 2= Yes)
11. **ChronicCond\_Depression:** Have a depression medical record? (1 = No; 2= Yes)
12. **ChronicCond\_Diabetes:** Have a diabetes medical record? (1 = No; 2= Yes)
13. **ChronicCond\_IschemicHeart:** Have an Ischemic Heart medical record? (1 = No; 2= Yes)
14. **ChronicCond\_Osteoporasis:** Have an osteoporosis medical record? (1 = No; 2= Yes)
15. **ChronicCond\_rheumatoidarthritis**: Have a rheumatoid arthritis medical record? (1 = No; 2= Yes)
16. **ChronicCond\_stroke**: Have a stroke medial record? (1 = No; 2= Yes)
17. **IPAnnualReimbursementAmt**: Total annual amount of Inpatient reimbursement per year
18. **IPAnnualDeductibleAmt**: Total annual amount of Inpatient deductible per year
19. **OPAnnualReimbursementAmt**: Total annual amount of Outpatient reimbursement per year
20. **OPAnnualDeductibleAmt**: Total annual amount of Inpatient deductible per year

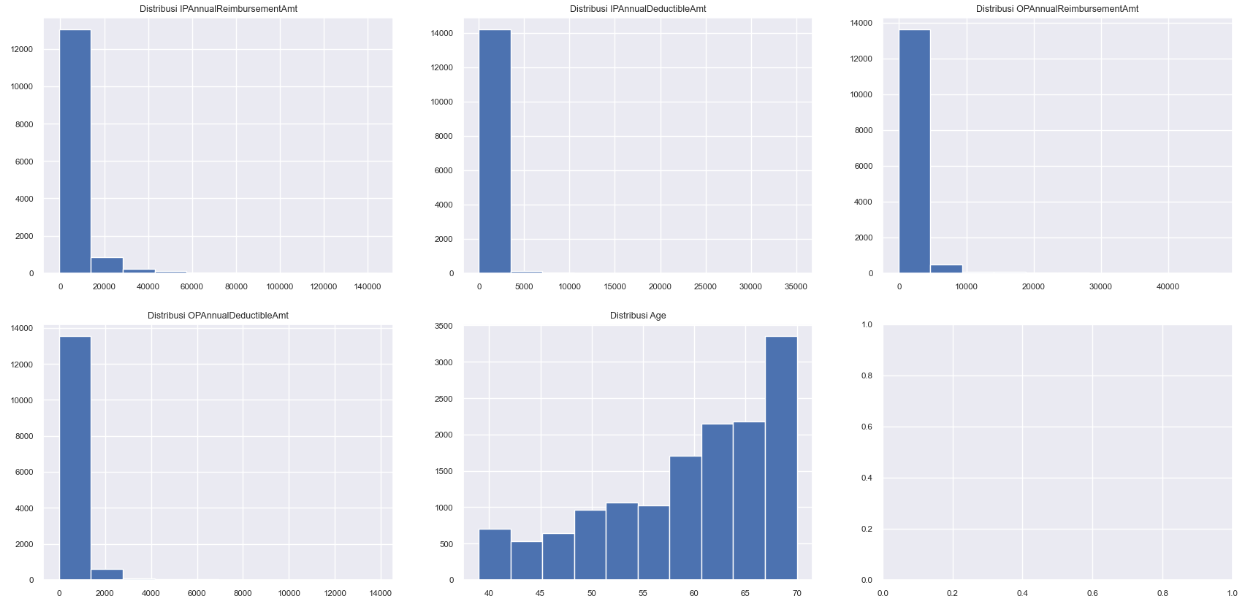
Missing value only found in Date of Death (**DOD**) columns, since the definition of the columns the missing value/columns can be ignored. Average maximum age limit for most term insurance plans is 65-70 years. However, term insurance can be buy even after the age of 65 years, albeit with a high premium amount. Dataset in this project will be limited to Age ≤ 70 years old, total data rows is reduced from **138556** to **14298.**



1. **Exploratory Analysis**

**3.1 Distribution of Inpatient-Outpatient Reimbursement and Deductible Amount**

Distribution of numerical data is relatively dominant on specific range with highly skewed distribution. Age of policy holder is dominantly claimed by > 60 years old members. The others variable such as Inpatient Annual Reimbursement amount is predominant in range of $0 - $20,000 and Outpatient Annual Reimbursement amount is predominant in range of $0 - $5,000. Inpatient Deductible amount shown by plot below predominant in range of $0 - $5,000 and Outpatient Deductible in range of $0 - $2,000. Since low deductible amount means larger annual premium, the policy holder in our data show that they are likely prefer to choose large annual premium payment which is good to insurer company. Inpatient is preferable to have larger amount rather than Outpatient since it requires to be admitted to a hospital or other care facility. However low deductible amount also shown that the policy holder is frequently claims the insurance.



**3.2 Proportion of All Variables**

The plot below shows the percentage of proportion from each variable. It is observed that Race have large difference of proportion between race, Renal Disease Indicator have quite large difference between person with positive renal disease (85.24%) and negative renal disease (14.76%). As well as person with both Cancer (No: 90.37%; Yes: 9.63%) and Stroke (No: 92.55%; Yes: 7.45%) which is most dominantly by person with no both disease in medical record. Bar plot shows almost all health condition variable has less person with Chronic Disease. However, there’s people with Heart Failure condition have large proportion 48.15% as well as people with Depression, Diabetes, and Ischemic Heart. The policy holder with these conditions should be warned to have low deductible amount since their condition might be turned them to frequently having outpatient treatment.

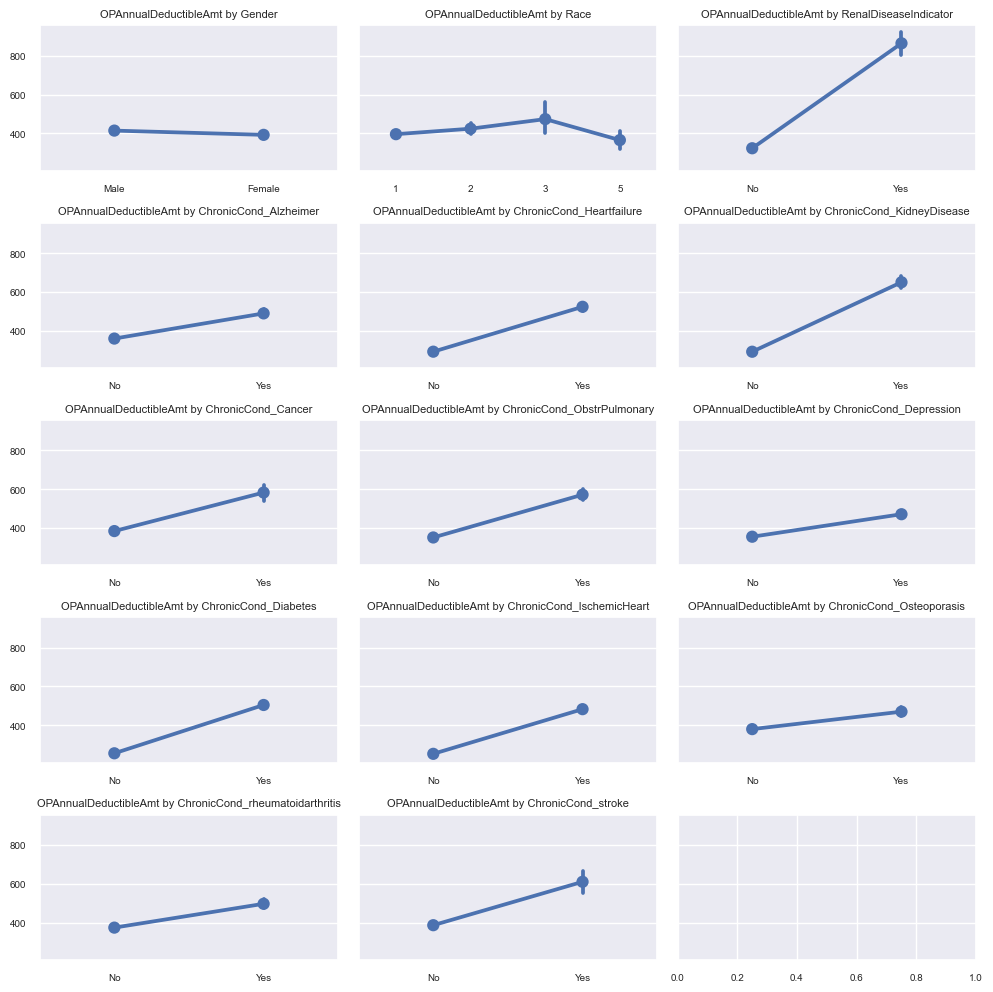


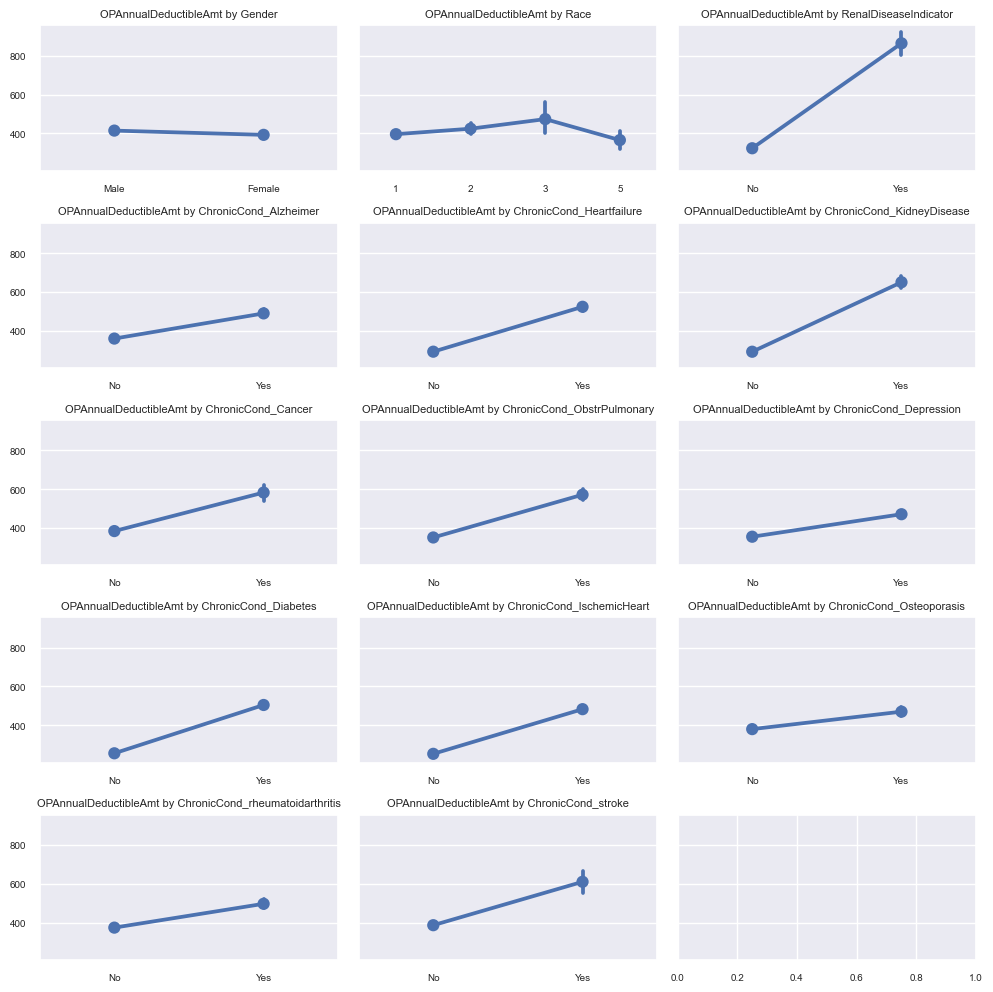
**3.3 Distribution of Variable toward Outpatient Deductible Amount**

Distribution of every variable toward OP Deductible amount shown by boxplot below. Mostly, all parameter shows OP deductible amount predominant distributed below $2,000 with much outlier. Variable Renal Disease Indicator, Heart Failure, Kidney Disease, Ischemic Heart show that there’s few people still prefer have high deductible amount despite of having severe potential to claim frequently. In reverse with variable Depression, Cancer, Osteoporosis, Rheumatoid arthritis, and Stroke show that people have outpatient deductible no more than $8,000. However, these are common diseases that often occur in hospital record and have more cases than the previous, hence these people having low deductible and high premium to prevent pay out more in deductible.



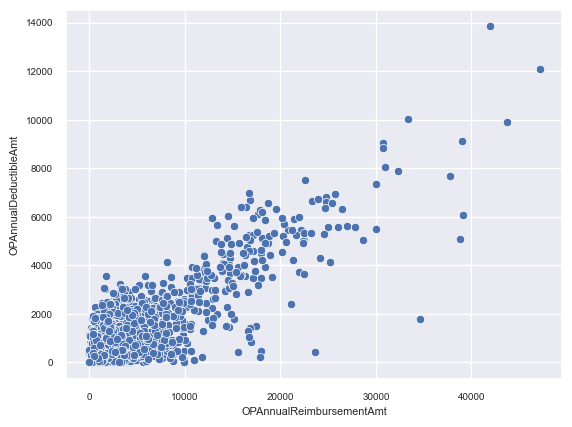
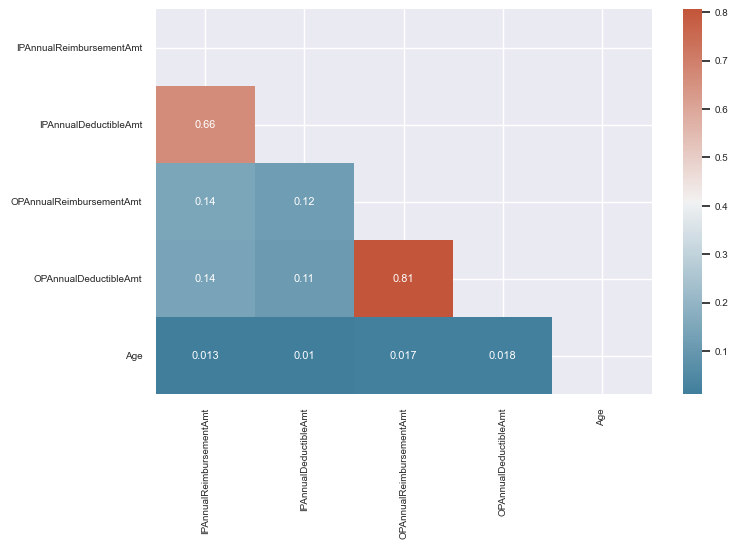
The picture below represents mean for each class of gender, race, and people with disease indication or not. Renal Disease Indicator, Kidney Disease and Stroke have more than $200 difference between each other.





**3.4** **Correlation between OP Deductible Amount and Numerical Variables**

Correlation show below is using Pearson correlation to calculate relationship between OP Deductible amount and the others numerical value. OP Reimbursement amount become the strongest positive correlation coefficient towards OP Deductible amount with +0.81, based on Pearson correlation, its relationship between these two variables moves in tandem or same direction. When OP Deductible increases, the other variable (OP Reimbursement) increases as well. This relationship shows that people with high OP Deductible amount and low premium preferable to have high reimbursement amount per year. Scatterplot below indicates that people with OP deductible under $4,000 mostly prefer to have under $20,000 reimbursement amount, and people with OP deductible over $4,000 doesn’t have reimbursement amount lower than $10,000.

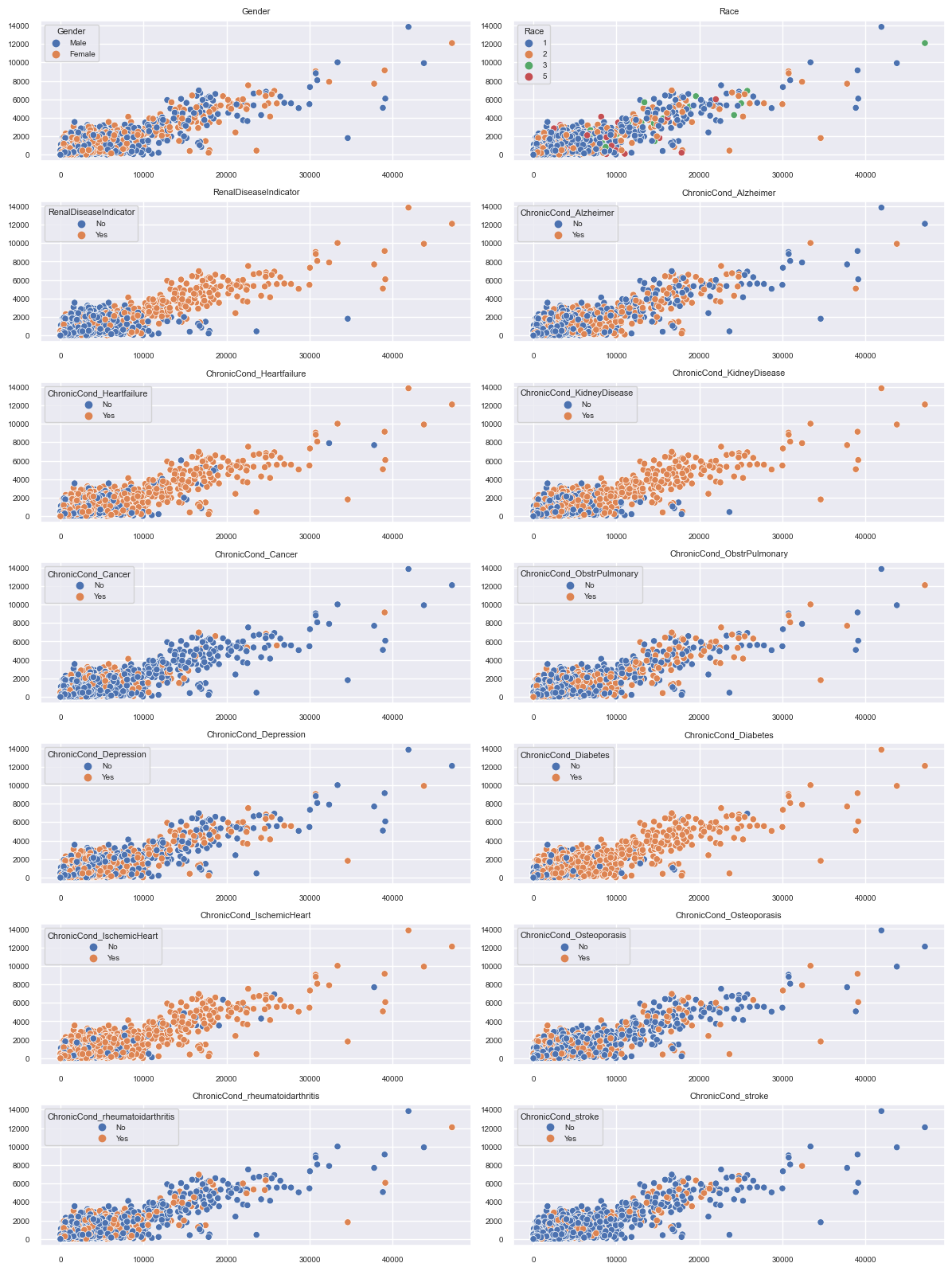


**3.5 OP Deductible against OP Reimbursement amount**

The people with positive renal disease indicator have predominant distribution with high OP Deductible amount and OP Annual Reimbursement rather than the people with negative renal disease indicator. It means people with positive renal disease having high deductible means that they might be not used to frequently check health or claim outpatient treatment in medical provider and can be indicates that they have low awareness to health condition

The people with positive kidney disease in medical record mostly prefer to have high OP deductible amount and high OP reimbursement rather than people with no kidney disease. It shown by distribution of scatterplot below. These peoples might be considered not to frequently claim outpatient treatment for their disease and more prefer to have low premium amount. It goes same with the other diseases.

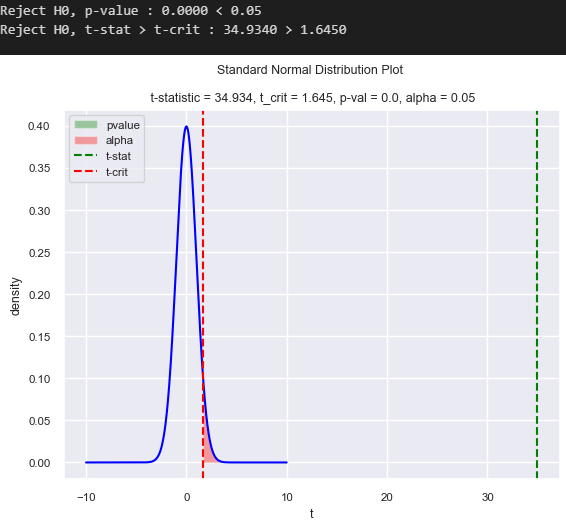
In few health diseases such as Cancer, Depression, Osteoporosis, Rheumatoid Arthritis and Stroke are dominated by people with negative result condition. This kind of disease has small mean margin between people with positive and not, but this distribution cannot be underline to conclude people with positive disease likely have low OP Deductible. Moreover, gender and race variable do not affect much in preferable of people to have low or high deductibility based on scatterplot below, mean category of these variable goes like others.



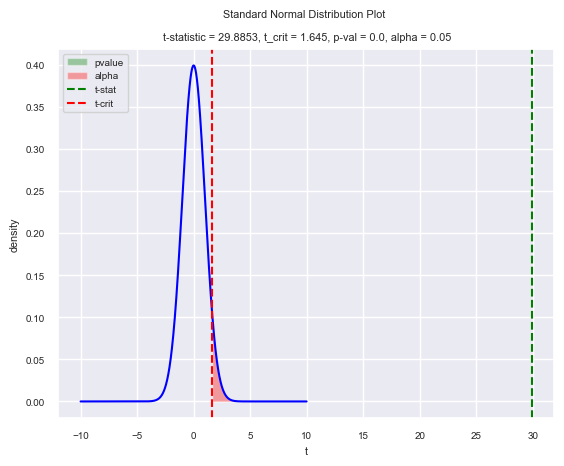
**4. Statistical Inference**

These data show the sample of several people (policy holder) with typical health condition, to inference the hypothesis we conclude above, statistical testing is used to show how these affects to population data. In statistical inference, we wish to make statements not merely about the subjects observed in a study but also, more importantly, about the larger population of subjects from which the study participants were drawn. Statistical inference method that are used in this project is statistical hypothesis testing. Statistical hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

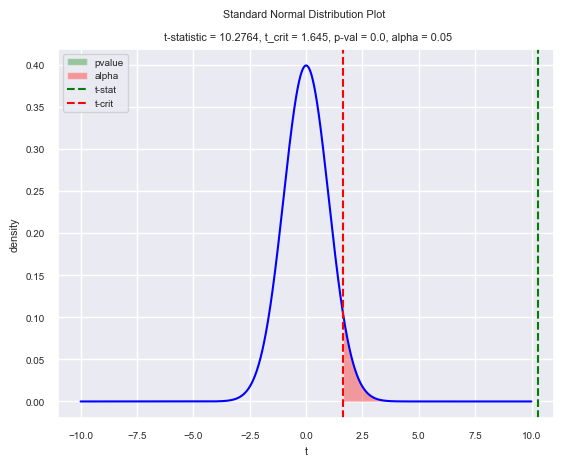
First hypothesis testing that we want to claim whether Outpatient Deductible Amount is higher for Policies with positive renal disease indicator. Null Hypothesis (H0) define with mean of the distribution underlying the sample of policies with positive renal disease is lower than mean of the distribution underlying the sample of policies with negative positive renal disease (Ho: μ ≤ μ0). The alternative hypothesis (H1) define that mean of distribution underlying the sample of policies with positive renal disease is greater than mean of distribution underlying the sample of policies with negative renal disease (H1: μ > μ0). We are using two-sample method with upper-tailed test since alternative hypothesis used is greater than. The method we are used is t-test statistical testing with two-sample with alpha 0.05. The result shows P-value is 0 and lower than alpha 0.05, t-statistic with value 34.94 is higher than t-critical in 1.645. Based on result, we can reject Null Hypothesis (H0), hence the OP Deductible amount of people with positive renal disease indicator is higher than negative one.



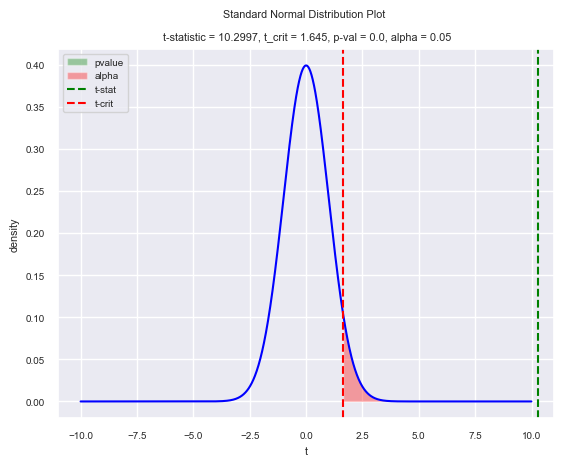
Next, the policies or people with kidney diseases have higher OP Deductible amount than the other one. Null Hypothesis (H0) define mean of the distribution underlying the sample with kidney diseases is lower than no kidney disease (H0: μ ≤ μ0) as well as the Alternative Hypothesis (H­1: μ > μ0). Test are used two sample t-statistic with upper tailed-test and alpha 0.05. The result shows P-value is nearly 0 and lower than alpha (pvalue < alpha), and t-statistic (29.89) is higher than t-critical (1.645), hence we reject Null Hypothesis (H0).



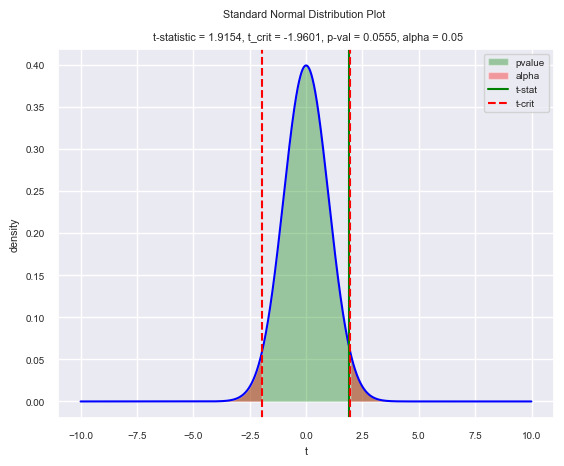
The Cancer disease scatterplot is dominated by people with no disease rather than the others. This disease has small mean margin between people with positive and not, but this distribution cannot be underline to conclude people with positive disease likely have low or high OP deductible amount. Performing t-test statistical of two sample between people with cancer disease and not, the assumption of this test is that people with cancer disease means are higher than the others. Null Hypothesis define (H0: μ ≤ μ0) as well as Alternative Hypothesis (H1: μ > μ0). The result show that P-value of 0.0 is lower than alpha and t-statistic (10.28) is greater than t-critical (1.645), hence we reject the Null Hypothesis.



It goes same with the other disease such as Stroke with t-test statistical of two sample with upper-tailed test. The result show p-value nearly 0.0 is lower than alpha and t-statistic (10.3) is greater than t-critical (1.645), hence we reject the Null Hypothesis.

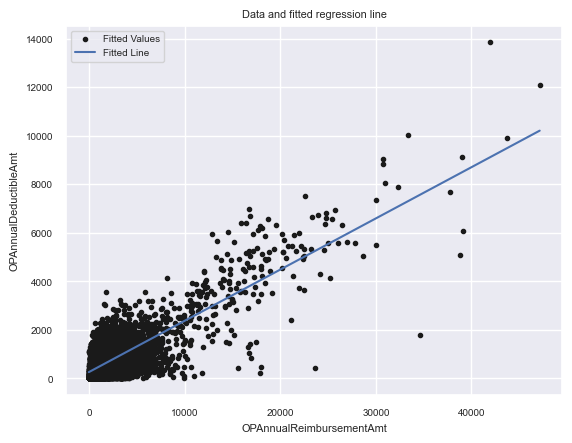


However, since Gender distribution between Male and Female can’t completely separate, we perform t-test statistical of two-sample to test whether OP Deductible amount for Male and Female is equal or not. Null Hypothesis (H0) define OP deductible amount for both is equal (H0: μ = μ0) and Alternative Hypothesis is (H1: μ ≠ μ0). The result show P-value (0.0555) is greater than alpha (0.05) and t-statistic (1.915) is between t-critical (±1.960), hence we failed to reject Null Hypothesis that indicates OP deductible amount between male and female is not equal but almost equal.



**5. Model Fitting**

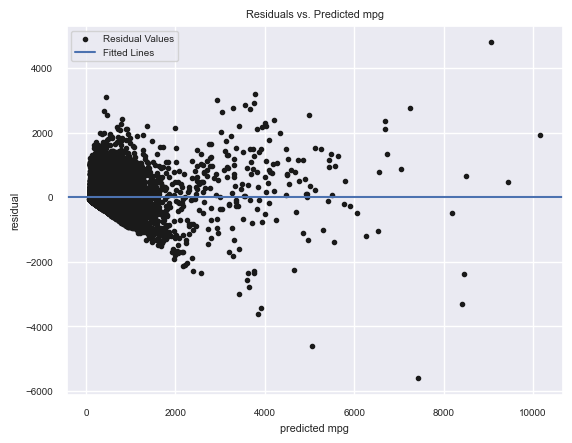
Based on Statistical Inference, we can drop unsignificant variable and use few variables such as OP Annual Reimbursement Amount, Renal Disease Indicator, Kidney Disease, Cancer, and Diabetes to build Linear Regression of OP Annual Deductible Amount. Firstly, we are used Cross Validation with 5 Folds to test the model regression performance toward new unfaced data. The average R2 score is 0.654 with the highest R2 score in Folds 2 with 0.673. This score is good performance in Linear Regression model performances; however, it should be good with the higher R2 score.



**OP Annual Deductible = 274.386 + 0.21OP Annual Reimbursement - 41.44KidneyDisease - 48.03Cancer – 33.94Diabetes - 87.51RenalDiseaseIndicator**

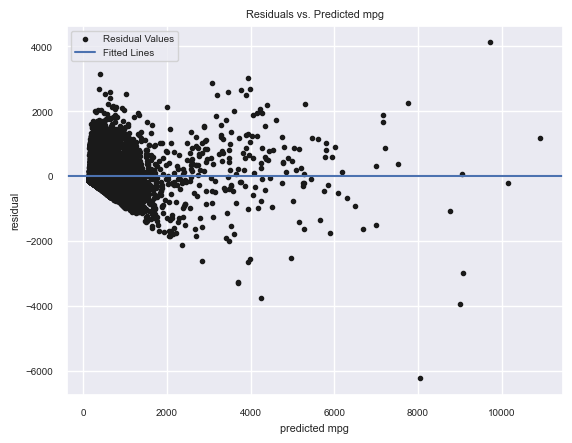
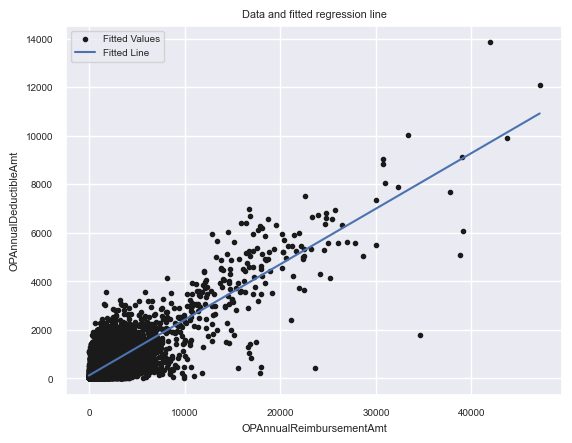
The average of OP Annual Deductible amount is $274.386 for people that has 0 OP Annual Reimbursement and zero Kidney Disease, Cancer, Diabetes and Renal Disease Indicator (people with no these disease). In these data comparing people with same OP Annual Reimbursement and did not have Kidney Disease, Cancer, and Diabetes but positive Renal Disease Indicator have predictive difference of $87.51. The predictive difference comparing two people that have same OP Reimbursement Amount, Kidney Disease, Cancer, and Renal Disease Indicator but differ 1 point in Diabetes is $33.94. The predictive difference comparing two people that have same OP Reimbursement Amount, Kidney Disease, Diabetes, and Renal Disease Indicator but differ 1 point in Cancer is $48.03. Comparing of two people that have same OP Annual Reimbursement, Cancer, Diabetes, and Renal Disease Indicator but differ 1 point in Kidney Disease have predictive difference in $41.44. Comparing of people with same indicator of Kidney Disease, Cancer, Diabetes, and Renal Disease Indicator but have differ 1 point in OP Annual Reimbursement amount is $0.21.

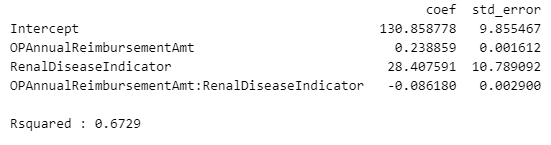
The residual plot below is obtained by reduce fitted values with the fitted lines values. The residual plot is does not show any pattern and accumulated nearly fitted lines of zero, hence this leads the model have rather good performance. However, there is few residual points that have high value that far from fitted lines indicates the Linear Regression model still need to improved.



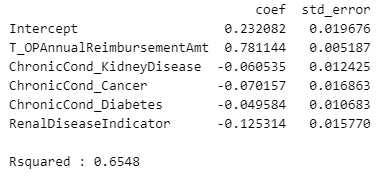
**OP Annual Deductible = 130.86 + 0.239OP Annual Reimbursement + 28.41Renal Disease Indicator – 0.086OP Annual Reimbursement \* Renal Disease Indicator**

The intercept represents the average OP Annual Deductible for people whose did negative Renal Disease Indicator (0) and zero OP Annual Reimbursement is $130.86. The estimated difference between OP Annual Deductible whose Renal Disease Indicator is 0 but differ 1 point in OP Annual Reimbursement is $0.24 and estimated difference between OP Annual Deductible whose OP Annual Reimbursement is 0 but differ 1 point in Renal Disease Indicator is $28.41. The coefficient on the interaction term, $-0.086 represents the difference in the slope for OP Annual Reimbursement, comparing people who did and did not have positive Renal Disease Indicator. The interaction linear regression of Renal Disease Indicator with OP Reimbursement amount has the higher R2 scored of 0.6729.





Since the scatterplot does not show any curve pattern the log transformation is not being applicable in this case, it goes same with the standardization transformation. R2 score achieved on standardization transform goes to 0.6548, non-significant increasing of R2 score.



**6. Limitation of Work**

While carrying out this work, the limitation we encounter are explained below:

1. Lack of relevant and correct source to gain variable regarding in separate OP Annual Deductible Amount category, basically OP Annual Deductible is accumulated under $10,000.
2. Lack of variation between variable, each variable show that positive Chronic Cond of disease always have higher OP Annual Deductible rather than the other one.

**7. Conclusion**

Kidney Disease, Cancer, Diabetes and Renal Disease Indicator variables has significant influence in consideration of OP Annual Deductible amount for many policies holder based on proportion, mean difference, and scatterplot of sample toward OP Annual Deductible amount. The others diseases, race and gender did not affect OP Annual Deductible amount predominantly. The policies holder with these diseases mostly prefers to have high OP Deductible amount and lower Annual Premium to insurer. These indicates the policies holder believed that they are not frequently having Outpatient Treatment, but keep preserve their health by bought low premium policy.