**ANALYSIS OF DEATH FACTORS AND SEVERAL CAUSES**

KOLIPAKA PREETHI

**ABSTRACT**

The primary or underlying cause of death is defined as that condition or injury (or circumstances of the injury) that initiated the train of morbid events leading directly to death. The question sometimes arises as to which of several existing conditions has caused death. The clinician may logically say that none of the diseases singly, in a specific patient, caused death, but rather the complex of conditions.

Analysis of multiple-cause-of-death will confirm prevalence rates of diabetes among racial or ethnic minority populations, demonstrate the impact of diabetes in association with other causes of death, and highlight variations of burden of disease among different racial or ethnic groups.

When a death is classified by a single underlying cause, information regarding the conditions that contributed to the death is lost. Multiple-cause data characterise the simultaneous or sequential occurrence of elements that may be involved more accurately. All of the morbid illnesses, diseases, and injuries listed on the death certificate are included in this data. The limits of evaluating mortality statistics in terms of a single cause of death have been known for several decades. Several such studies have been undertaken in a number of nations as a result of developments in automated coding that make multiple-cause analyses possible.

**DATA COLLECTION**

This analysis is based on the datasets consisting of 6469 records entirely and the subset of this data is extracted from the entity called world and several regressions and analysis are performed over the extracted dataset and the data set also contains numerous causes of death.

**APPROACH**

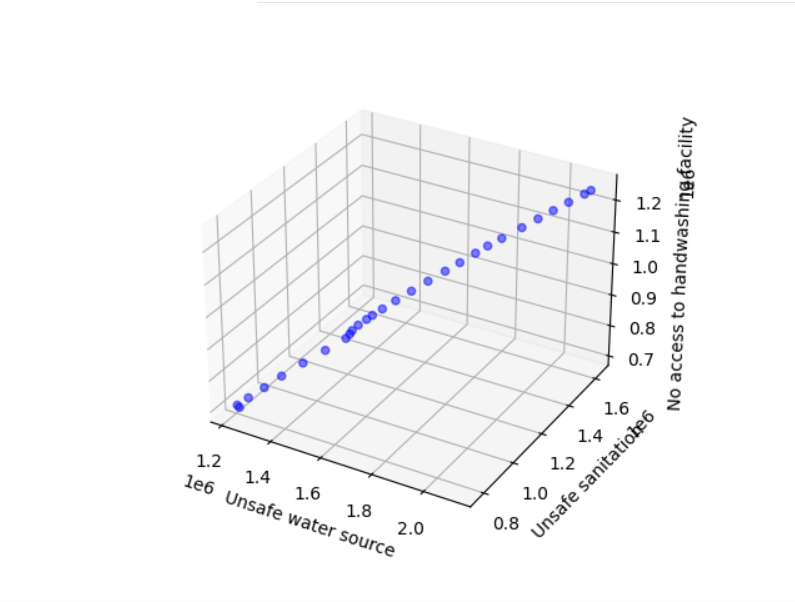
* The various python libraries such as *Numpy*, *pandas*, *Matplotlib* are used for the purpose of mathematical calculations, extraction of data and visualization respectively.
* The extracted dataset was found to have *null values* which are then effectively handled to achieve accurate analysis rather than removing them from the dataset.
* Different features measured in different scales will be perplexing while comparison. Hence, it is scale down using *StandardScaler()* function in *sklearn* library, which can convert all the values to a common format and makes analysis easier

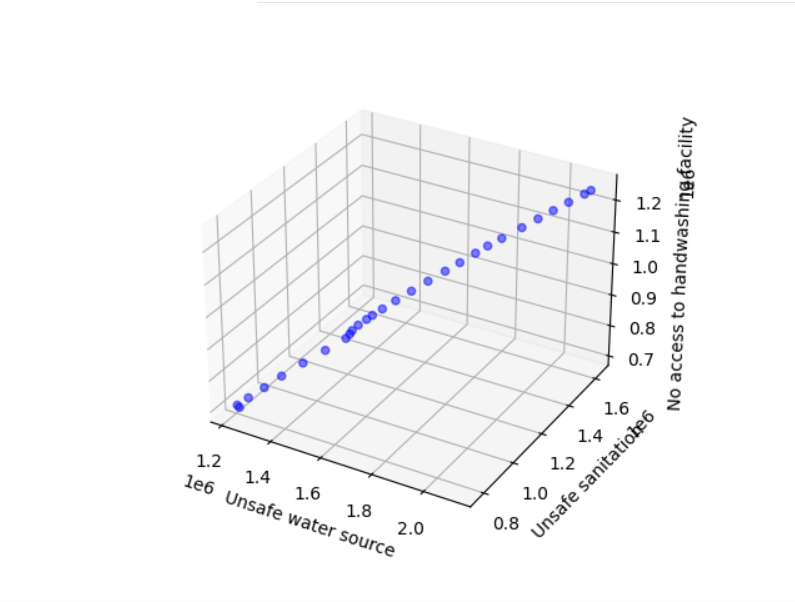
1. **MULTIPLE REGRESSION ANALYSIS**

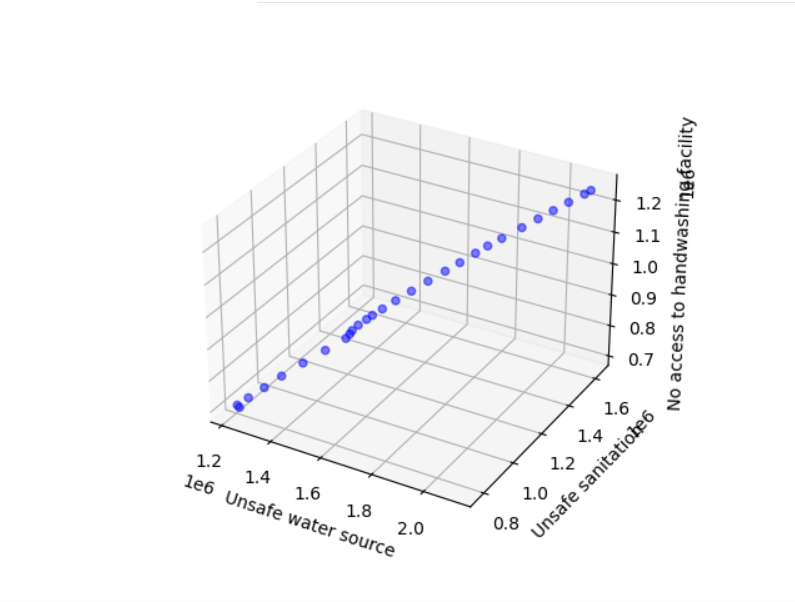
* **Analysis of Unsafe Sanitization and Water Source With respect to No Proper Access to Hand Washing Facility.**

PURPOSE

* + In Order to know the contribution of Unsafe sanitization and Unsafe water source over the dependant variable and which contributes more to no proper access to hand washing facility.







Prediction: [89153.05851786]

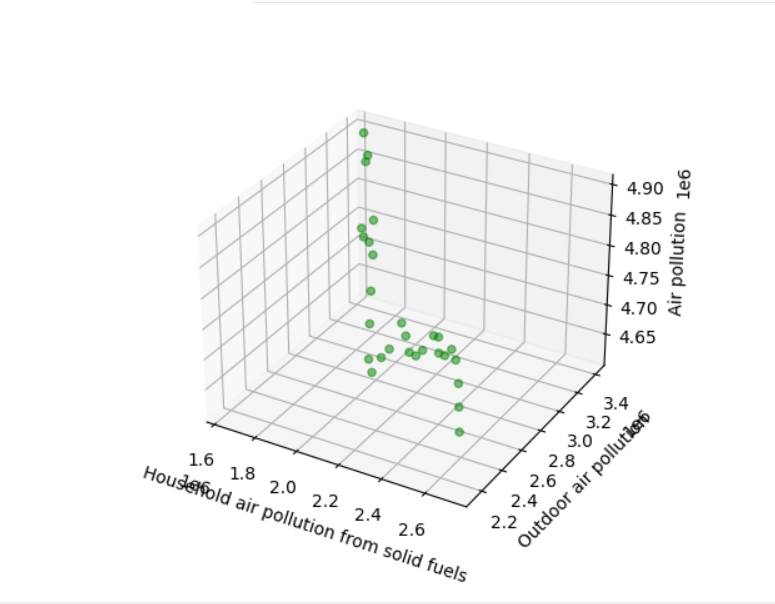
Coefficients: [0.31664531 0.30131765]

INFERENCE:

* From the given data, considering only world entity, we can observe that unsafe water resource and unsafe sanitization are two independent variables and No access to hand washing facility can be considered as dependent variable. The Coefficient for Unsafe Sanitization is 0.31664531 which is higher than that of coefficient of unsafe water resource i.e., 0.30131765
* The Contribution of unsafe Sanitization for the dependant variable i.e. No proper access to hand washing facility is 0.3166 which is more than the contribution of unsafe water resource. We can say that for every one unit increase over the independent variable (Unsafe sanitization) can result in 0.3166 rise for the No proper access to hand washing facility.
* We can say that Proper utilization of water resources and proper sanitization can play a specific role in developing the entire world.
* **Analysis of Air pollution with respect to household air pollution from solid fuels and outdoor air pollution.**

PURPOSE

* + In Order to know the contribution of Outdoor air pollution and Household air pollution over the dependant variable and which contributes more to Air pollution.



Prediction: [190991.12614171]

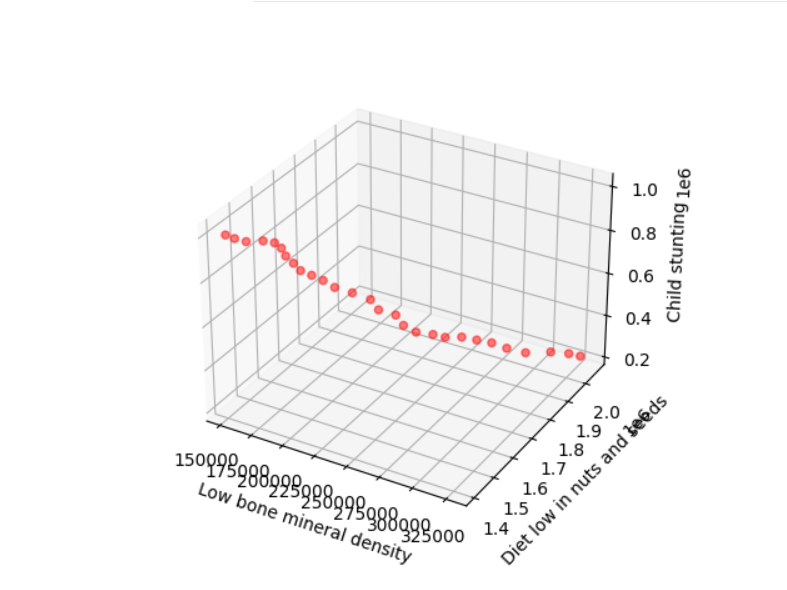
Coefficients: [0.92075656 0.94338962]

INFERENCE:

* From the given data, considering only world entity, we can observe that Household air pollution and outdoor air pollution are two independent variables and Air Pollution can be considered as dependent variable.
* The Coefficient for Household air pollution from solid fuels is 0.92075656 which is lower than that of coefficient of Outdoor air Pollution i.e., 0.94338962.
* We can say that for every one unit rise in Outdoor air pollution can cause 0.94338962 units rise in air pollution and for every unit rise in Household air pollution from solid fuels can cause 0.92075 units rise the air pollution over the world.
* **Analysis of Child Stunting and causes of it such as Diet low in nuts and seeds and Low Bone Mineral Density.**

PURPOSE

* + In Order to know the contribution of Low Bone Mineral Density and Low diet in nuts and seeds over the dependant variable and which contributes more to Child Stunting.



Prediction: [1508384.25329867]

Coefficients: [-4.78247808 0.07025614]

INFERENCE:

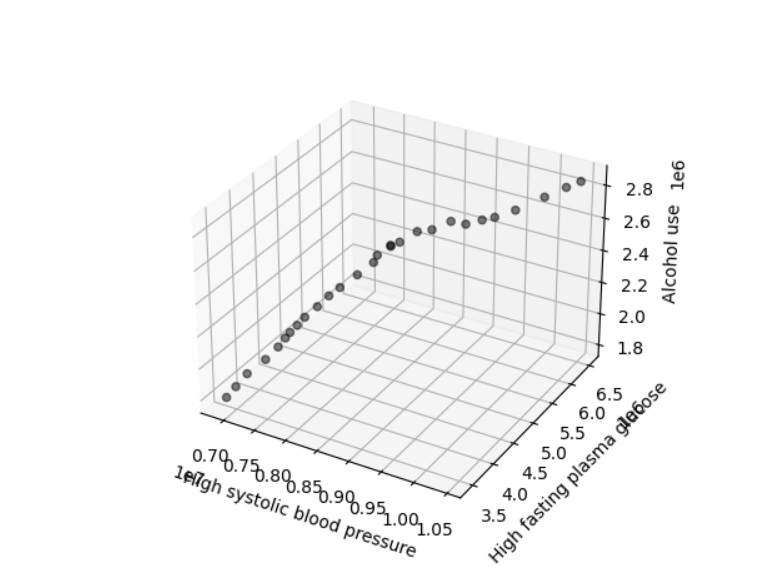
* From the given data, considering only world entity, we can observe that Low bone mineral density and diet low in nuts and seeds are two independent variables and child stunting can be considered as dependent variable.
* The Coefficient for Diet low in nuts and seeds is 0.07025614 higher than the coefficient of low bone mineral density -4.78247808.
* We can say that the contribution of diet low in nuts and seeds is more in child stunting and for every unit increase in diet low of nuts and seeds can cause 0.070 units increase in the contribution of child stunting.
* We can also say that proper intake of nuts and seeds can have adequate growth of height of a child.
* **Analysis of alcohol usage and causes of it such as Blood Pressure and seeds and High Fasting Plasma Glucose.**

PURPOSE

* + In Order to know the contribution of Low Bone Mineral Density and Low diet in nuts and seeds over the dependant variable and which contributes more to Child Stunting.

Prediction: [1859815.13578811]

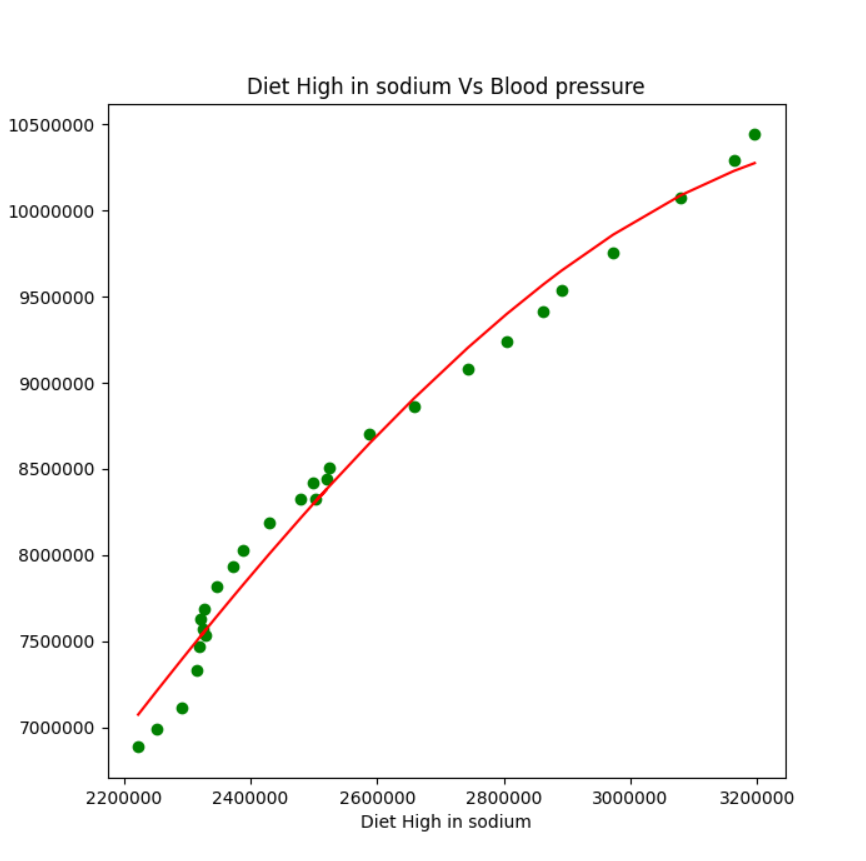
Coefficients: [-0.28709037 0.61752529]



INFERENCE:

* From the given data and graph, we can say that high fasting plasma glucose contribution is more i.e for a unit increase in High fasting plasma glucose can cause 0.6175 units rise in alcoholic level.
* We can also say that more glucose contribution is involved in alcohol usage.
* And Increase in glucose levels as well as alcoholic consumption can cause High systolic blood pressure.

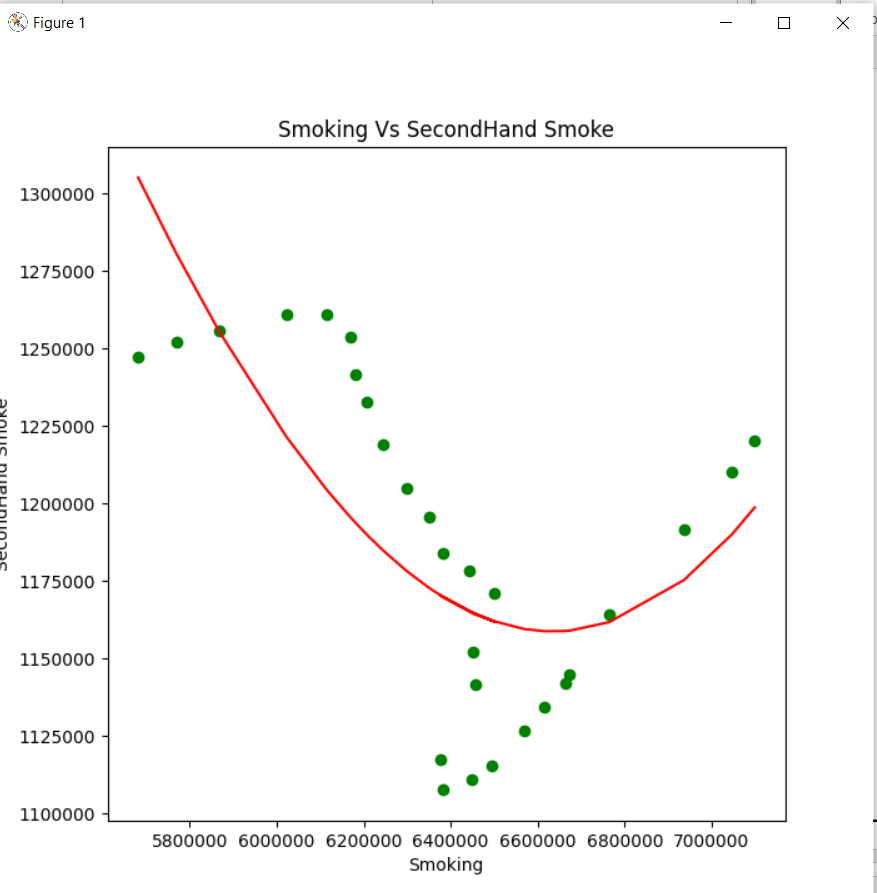
1. **POLYNOMIAL and LINEAR REGRESSION ANALYSIS**
2. **Diet High in Sodium Vs Blood Pressure**



INFERENCE:

* Polynomial Regression technique is applied on number of deaths caused due to Diet High in Sodium and High fasting Blood Pressure.
* The r2 score obtained is 0.9927750838902037 which is almost close to 1 . This implies deaths caused due to High fasting Blood Pressure is more depending on Diet high in Sodium.

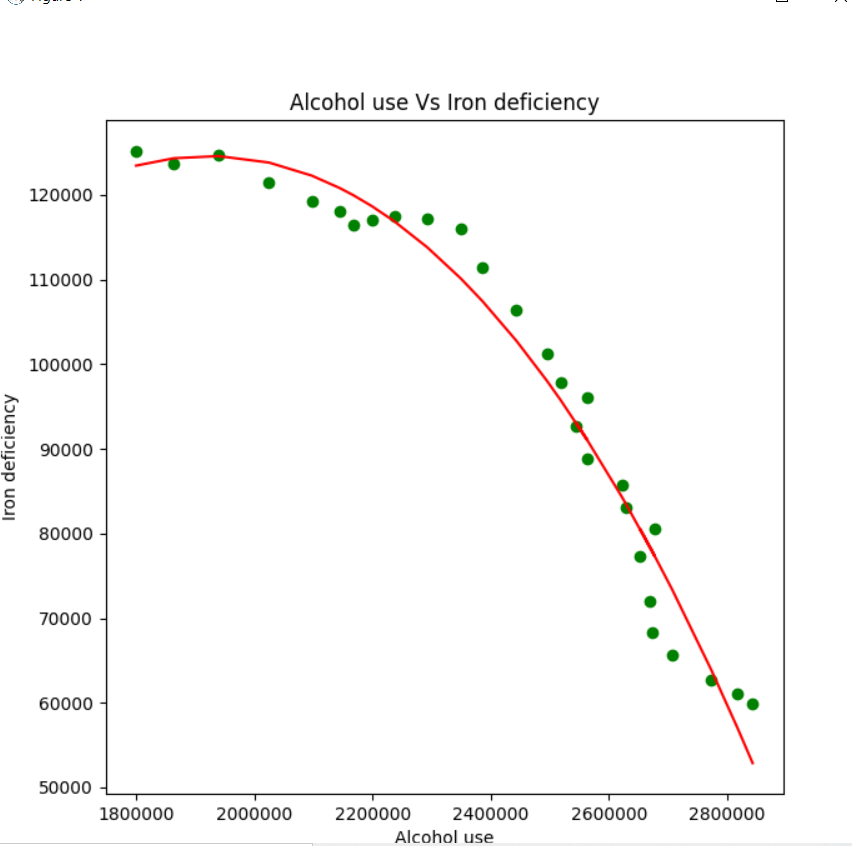
1. **Smoking Vs Second-hand Smoke**



INFERENCE:

* Polynomial Regression technique is applied on number of deaths caused due to smoking and second-hand smoking.
* The r2 score obtained is0.7466874302810738which is near to close to 1 . This implies deaths caused due to Smoking also depends on the deaths caused due to Second-hand Smoke
* To prevent this pupil need to think before smoking as it not only effects them but also their surroundings.

1. **Alcohol Usage Vs Iron Deficiency**

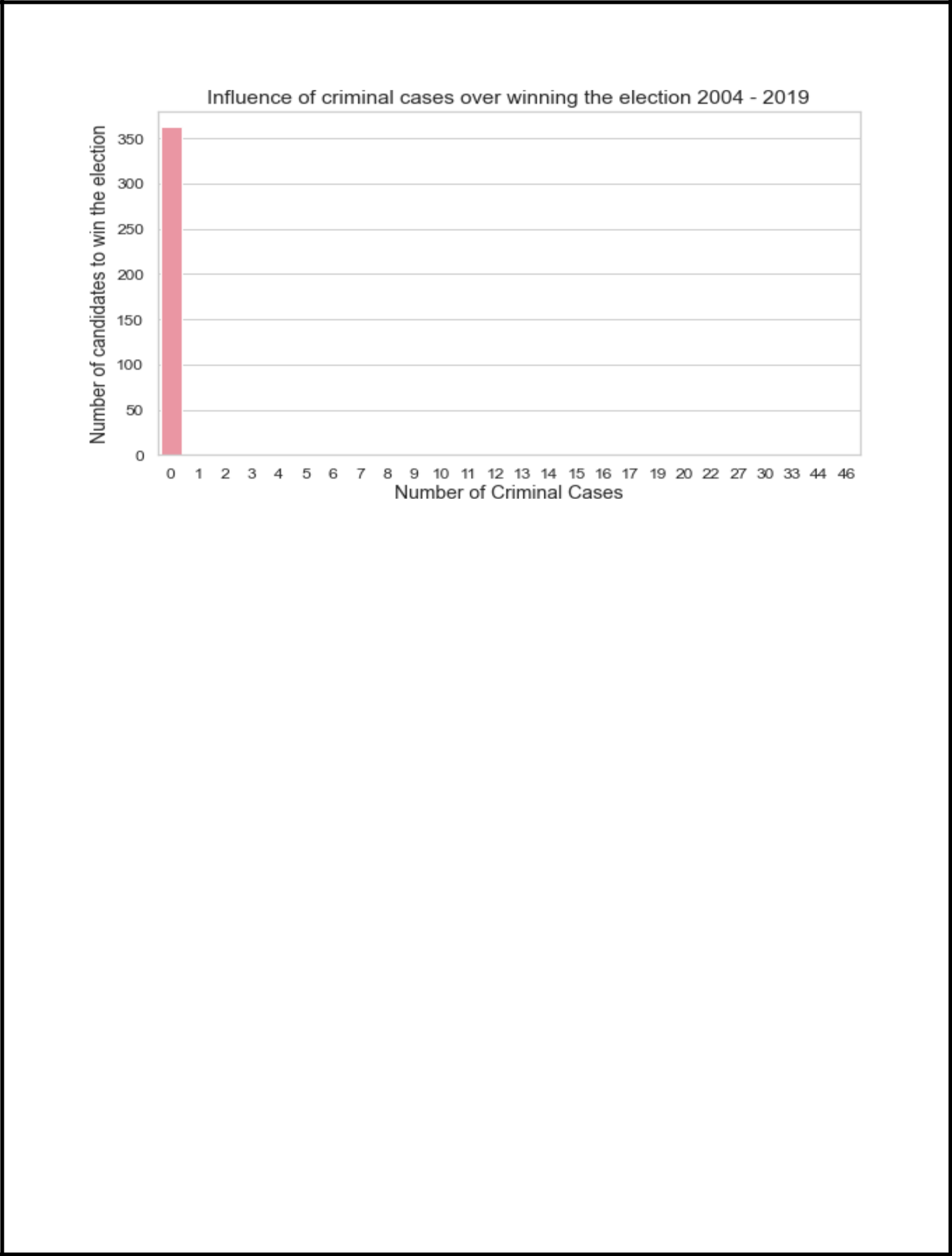


INFERENCE:

* After Scattering points on the graph , the Polynomial Regression is applied on number of deaths caused due to Alcohol consumption and Iron deficiency.
* The r2 score obtained is 0.9668266085066587which is almost close to 1. This implies deaths caused due to Alcohol usage is more depending on Iron deficiency.
* The Most deaths caused because of driving after consuming alcohol.
* In-order to avoid it drunk and drive tests were introduced.

**CONCLUSION**:

* Hence We can conclude that Deaths are caused with various factors .
* Each and Every Factor has its contribution towards death.
* Maintaining Proper diet and Following daily physical activity , avoiding alcohol consumption and smoking can lead to proper heath condition which maintains us to be active in our daily life.

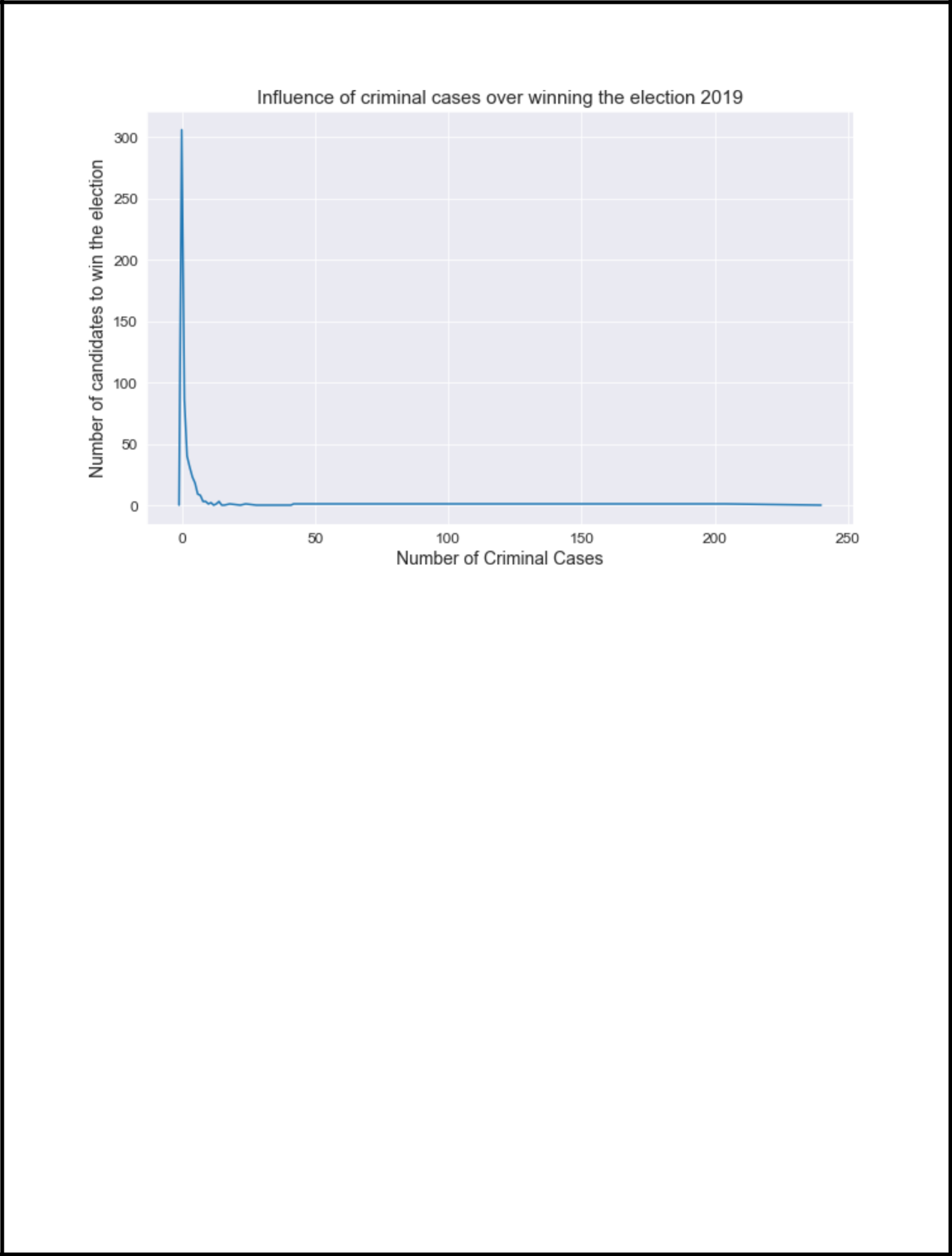


* It is commendable that, during the period from 2004 to 2019, the candidates with zero criminal cases have won the election.
* This proves the dependence of winning the election over zero criminal cases.

ANOVA Test on candidates in the year 2019

* Null hypothesis: The criminal cases of candidates has no significance in winning the election in 2019
* Level of significance: 0.05
* Since the P value is less than Level of significance (0.021985 < 0.05), we reject the null hypothesis. Thus, criminal cases of candidates have significance over winning the election in 2019.





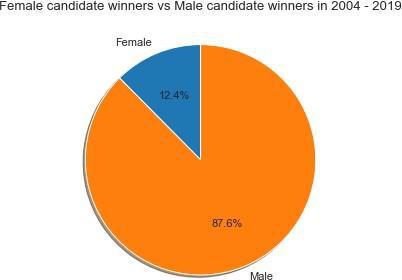
* From the above plot, it is evident that the winning candidates in 2019 do possess criminal cases

INFERENCE

* + From the above tests and plots we can infer that the number of criminals winning the election is increasing from the year 2019 and it is in the hands of public to reduce this rate.
  + So, a candidate to win an election most probably should have no criminal cases

1. **Influence of Female candidates in the outcome of the election result** ANOVA Test on Gender and winner of the election between 2004 – 2019
   * Null hypothesis: The gender of candidates has no significance in winning the election
   * Level of significance: 0.05
   * Since the P value is less than Level of significance (1.43e-08 < 0.05), we will reject the null hypothesis. Thus, gender of candidates has significance over winning the election during 2004 - 2019.

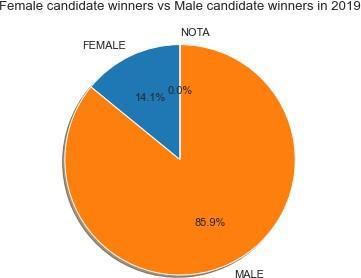




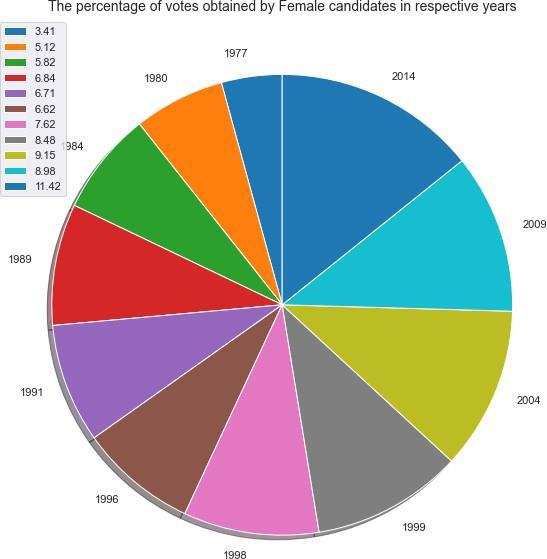
* It is evident that 45 out of 451 Female candidates have won in the election during 2004 – 2019

ANOVA Test on Gender and winner of the election in 2019

* Null hypothesis: The gender of candidates has no significance in winning the election
* Level of significance: 0.05
* Since the P value is less than Level of significance (5.39e-20< 0.05), we reject the null hypothesis. Thus, gender of candidates has significance over winning the election in 2019.



* In 2019 election, 76 out of 258 Female candidates have won the election The percent of votes obtained in election by the female candidate from 1977 – 2015



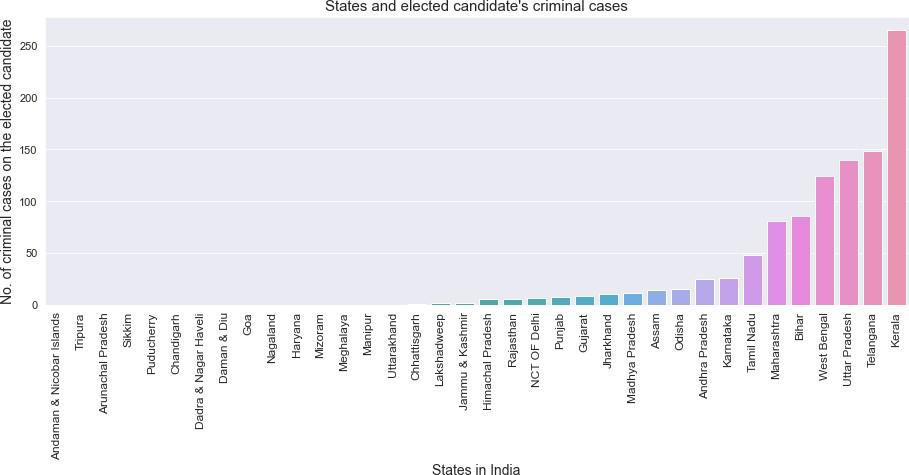
INFERENCE

* From the above tests and plots, we can infer that the number of female candidates participated in the election is comparatively lesser than that of male candidates.
* Although the female candidates are less in number, the rate of winning for a female candidate is higher than that of men.
* It is also obvious that the percentage of votes being secured by a female candidate is gradually increasing over the years from 1977 – 2014.
* Thus, the chances of winning an election by a female candidate is more than that of a male candidate.

1. **The state with highest number of criminals elected**

PURPOSE

* + The state with highest number of criminals elected is at high risk in both development and growth.



INFERENCE

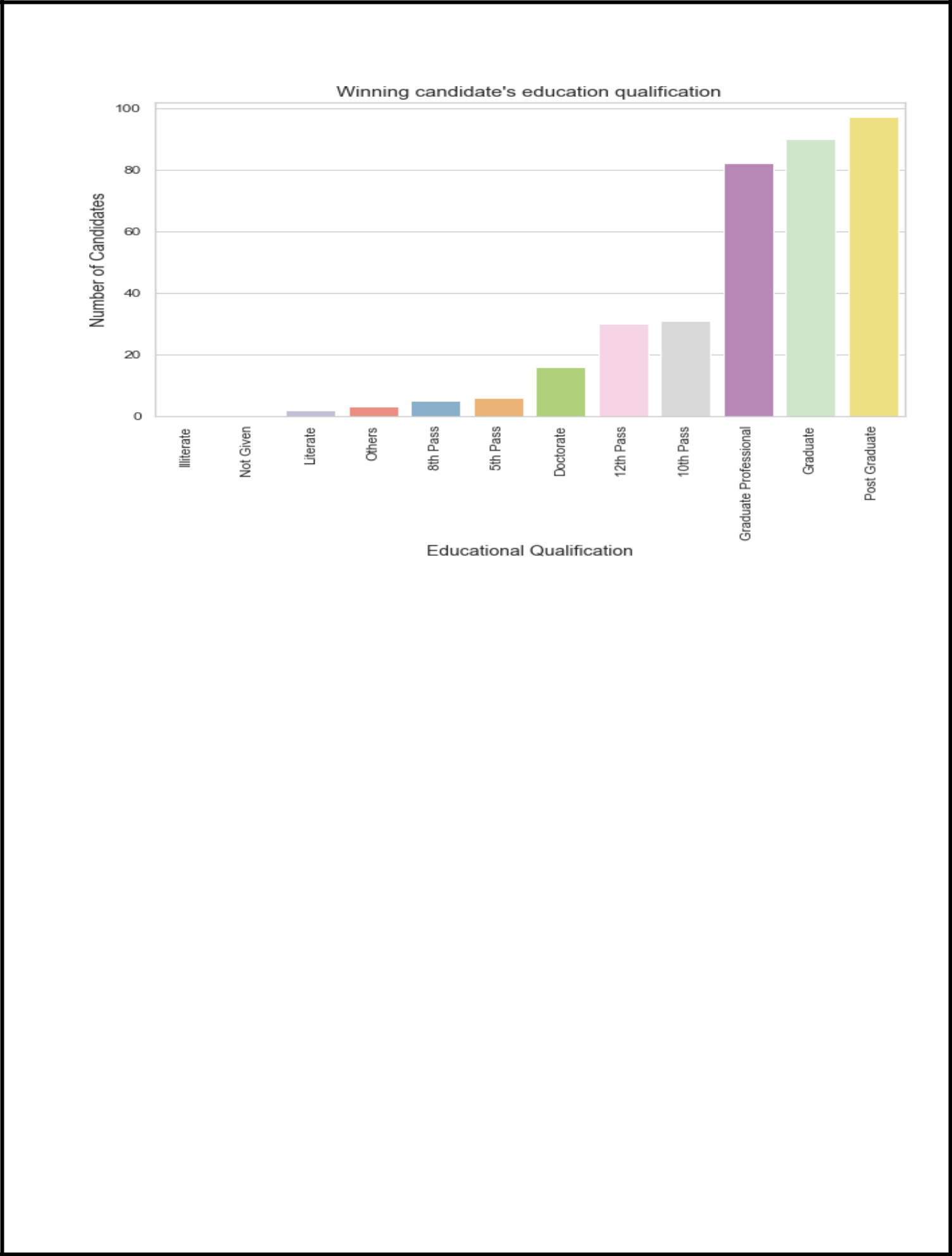
* + From the above analysis, it is crystal clear that Kerala has highest number of criminals elected.
  + This should be considered by the public and it can also pave a path for new candidates in the election with no criminal cases.

1. **Influence of educational qualification over winning the election**

ANOVA TEST on winner and educational qualification of a candidate between 2004-2019

* + Null hypothesis: The educational qualification of candidates has no significance in winning the election
  + Level of significance: 0.05
  + Since the P value is less than Level of significance (8.74e-49 < 0.05), we will reject the null hypothesis. Thus, educational qualification of candidates has significance over winning the election during 2004 - 2019.



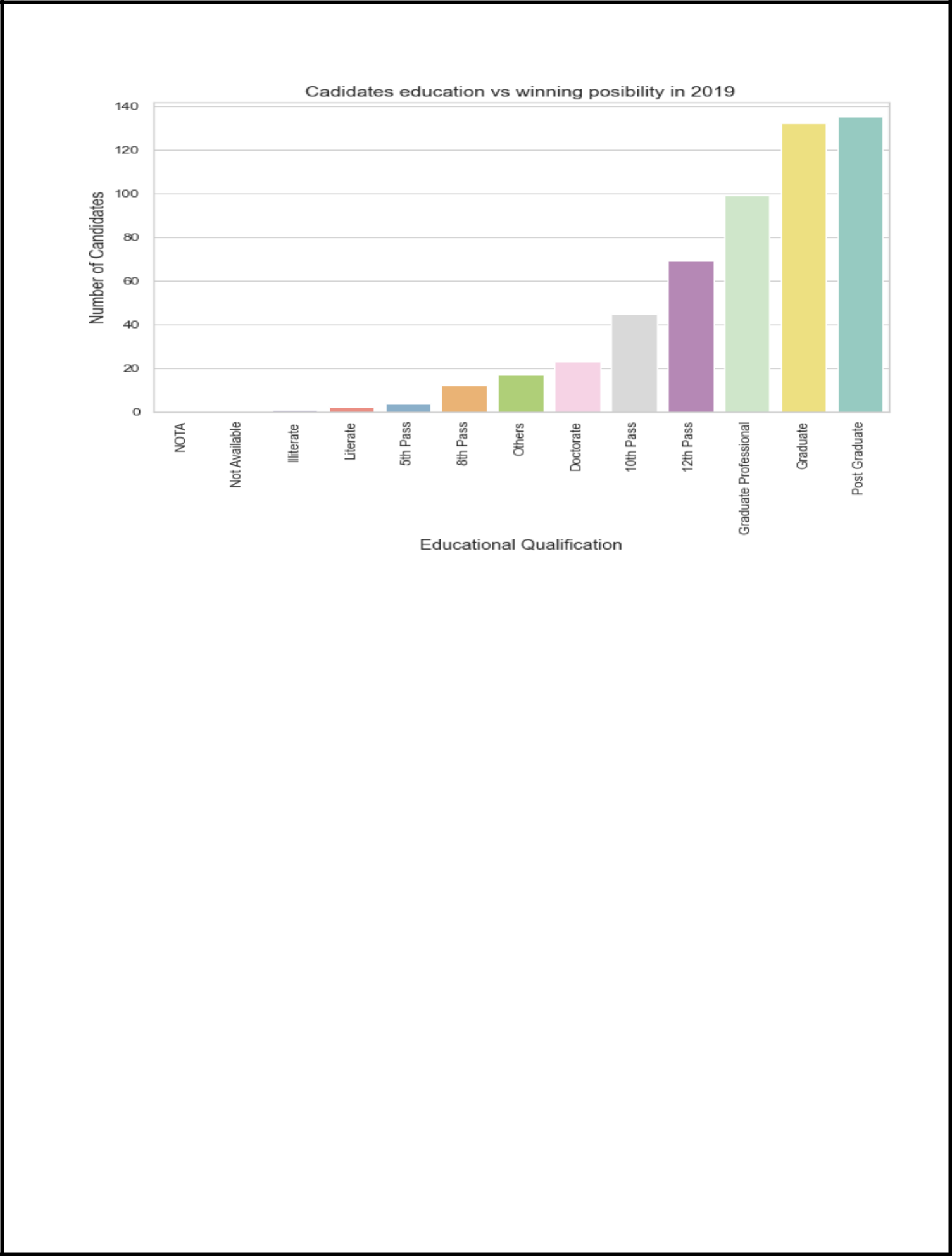


* From the above bar plot, we can deduce that illiterate doesn’t win any election, higher the educational qualification higher is the chances of winning an election.

ANOVA TEST on winner and educational qualification of a candidate in 2019

* Null hypothesis: The educational qualification of candidates has no significance in winning the election in 2019
* Level of significance: 0.05
* Since the P value is less than Level of significance (6.86e-20 < 0.05), we will reject the null hypothesis. Thus, educational qualification of candidates has significance over winning the election in 2019





**CONCLUSION**

* From the election analysis over the years between 2004 to 2019, we can arrive at the following insights
* The probability of a candidate with criminal cases to win an election is too low
* The number of female candidates is low but the rate of winning of a female candidate is higher than that of men
* Kerala is the state with highest number of criminals elected
* The probability of an educated candidate (most probably post graduate) to win an election is high.

**REFERENCE**

* Dataset for Election Results of 2014 and 2019:

<https://www.kaggle.com/ankit2106/indian-general-election-2019-candidate-wise-data>

* Dataset for candidate details during 2004 – 2019:

<https://www.kaggle.com/themlphdstudent/lok-sabha-election-candidate-list-2004-to-2019>

* Dataset for overall election results from 1977 – 2015:

<https://www.kaggle.com/awadhi123/indian-election-dataset?select=indian-national-level-election.csv>

* For code and plots

[https://github.com/Yashwanthra/election\_analysis/blob/main/Election\_data\_a](https://github.com/Yashwanthra/election_analysis/blob/main/Election_data_analysis.ipynb) [nalysis.ipynb](https://github.com/Yashwanthra/election_analysis/blob/main/Election_data_analysis.ipynb)