# WILDFIRES: THE EFFECTS ON SOUTHERN CALIFORNIA

A Thesis

Submitted to the Graduate Faculty of the

National University, Department of Engineering and Computing

in partial fulfillment of the requirements for the degree of

Masters of Science in Data Science

Prepared By:

Aakriti Sinha

Adam LeBlanc

Aditi Bhujbal

Anna Mkrtchyan

National University

November 20, 2022

**MASTERS THESIS APPROVAL FORM**

We certify that we have read the project of Aakriti Sinha, Adam LeBlanc, Aditi Bhujbal, and Anna Mkrtchyan entitled WILDFIRES: THE EFFECTS ON SOUTHERN CALIFORNIA and that, in our opinion, it is satisfactory in scope and quality as the thesis for the degree of Master of Science in Data Science at National University.

Approved:

Jedediah Baker Ph.D., Capstone Project Sponsor — Date
Adjunct Professor, Department of Engineering and Computing
National University

Barbara Lauridsen, Ph.D., Capstone Project Advisor — Date
Professor, Department of Engineering and Computing
National University

Aeron Zentner D.B.A., Capstone Project Instructor — Date
Adjunct Professor, Department of Engineering and Computing
National University

**Acknowledgement**

Members of A-Team would like to express our deep and sincere gratitude to the committee: Dr. Jedediah Baker (Sponsor), Dr. Barbara Lauridsen (Advisor), Dr. Aeron Zentner (Project Instructor), for their continued support, encouragement, and for providing their guidance throughout this research. We offer genuine appreciation for the learning opportunities provided by National University. We would also like to thank our classmates who contributed their valuable and timely feedback on our thesis paper. A heartfelt thanks also goes to the librarian of National University: Benita Ghura, who helped us find the relevant literature; and to the writing center tutor: Audrey Lapointe, who guided us in the writing process.

Finally, to our families - thank you for allowing each one of us the time away to work on this project, while you managed everything else without any help. The encouragement from everyone when the times got rough are much appreciated and duly noted.

Once again, thanks to all who helped us during this journey!

## Abstract

The purpose of the research project was to develop and test predictive models that make accurate statistical predictions of the wildfire frequency and severity in the Southern California region. From 2011 to 2022, the annual acreage destruction by wildfires increased by approximately threefold. The research focused on identifying correlated variables leading to the beneficial environment for these wildfires. The data was collected from existing data sources: CalFire, U.S. Carbon Monitor, Energy Information Administration, National Oceanic and Atmospheric Administration, U.S. Census Bureau, Federal Reserve Economic Data, and U. S. Department of Agriculture. A data mart was created to store and manipulate the data. The research confirms that explanatory variables such as weather, carbon emissions, and populations are linked to wildfires in Southern California, and the prediction models have generated results that helped accept the research hypothesis.



***Keywords:*** Wildfires, wildfire severity, wildfire frequency, southern California, carbon emission, weather, climate change, population, WUI, Wildland-Urban-Interface, XGBoost, data mart

## Table of Contents


Acknowledgement ..... iii
Abstract ..... iv
List of Tables ..... vii
List of Illustrations ..... viii
Chapter 1 - Introduction ..... 1
Background ..... 1
Problem Statement ..... 2
Research Hypotheses ..... 4
Objectives ..... 7
Carbon Emission ..... 8
Weather ..... 9
Population ..... 10
Limitations of the Study ..... 12
Carbon Emission ..... 12
Weather ..... 12
Population ..... 13
Wildfire ..... 13
Definition of Terms ..... 14
Summary ..... 15
Chapter 2 - Literature Review ..... 16
Wildfire Frequency and Severity ..... 16
Carbon Emission and Wildfires ..... 21
Effects of Weather on Southern California Wildfires ..... 24
Population growth in Wildland-Urban Interface area ..... 28
Uniqueness about Research ..... 33
Chapter 3 - Methodology ..... 34
Data Sources and Collection Methods ..... 34
Wildfire Data ..... 34
Carbon Emissions Data ..... 36
Weather Data ..... 38

Population Data ..... 39
Data Cleaning ..... 43
Data Mart ..... 44
Analysis Methods ..... 48
Input Variable Selection for the Models ..... 48
Multiple Linear Regression Model ..... 49
Extreme Gradient Boosting (XGBoost) ..... 50
Spatial Data Analysis ..... 50
Summary of Methodology ..... 51
Chapter 4 - Results and Analysis ..... 52
Exploratory Data Analysis on individual datasets ..... 52
EDA on Variables from Wildfire Dataset ..... 52
Carbon Emissions And Fuel Consumption Variables ..... 58
Precipitation ..... 65
Maximum Temperature ..... 67
Minimum Temperature ..... 69
TotalPop And Income Variables ..... 71
Area_SquareKm, Housing_number, and Population_count ..... 76
Exploratory Data Analysis on Master Data ..... 78
Initial exploratory analysis of selected outcome and explanatory variables ..... 79
Prediction Models ..... 92
Regression Model ..... 92
XGBoost Prediction Model for Wildfire Frequency ..... 97
XGBoost Prediction Model for Wildfire Severity ..... 100
Chapter 5 - Conclusions and Recommendations ..... 104
Conclusion ..... 104
Limitations ..... 106
Implications for Practitioners ..... 107
Recommendations and Future Scope ..... 107
Summary ..... 108
References ..... 109

## List of Tables


| | | |
|---|---|---|
| Table 1 | Definition of Terms……... | 14 |
| Table 2 | Multivariate Analysis of Variance for Acreage Burned…… | 58 |
| Table 3 | Summary statistics of variables Metric_Tonnes_of_CO2_Emissions And Total_Fuel_Consumption_(MMBtu)……. | 60 |
| Table 4 | Summary statistics of variables Metric_Tonnes_of_CO2_Emissions And Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable…….. | 60 |
| Table 5 | PROC FREQ of all variables within the Carbon Emission dataset……... | 61 |
| Table 6 | Descriptive Statistics of Precipitation…… | 66 |
| Table 7 | Descriptive Statistics of Maximum Temperature…….. | 68 |
| Table 8 | Descriptive Statistics of Minimum Temperature……... | 70 |
| Table 9 | Summary statistics of variables TotalPop and Income…….. | 73 |
| Table 10 | Summary statistics of TotalPop using County as classification variable…….. | 73 |
| Table 11 | Summary statistics of Income using County as classification variable……. | 74 |
| Table 12 | Summary statistics of Area_SquareKm, Housing_number, and Population_count…….. | 76 |
| Table 13 | Summary statistics of Area_SquareKm using WUI_Category as classification variable.…… | 78 |
| Table 14 | Summary statistics of Housing_number using WUI_Category as classification variable…… | 78 |
| Table 15 | Summary statistics of Population_count using WUI_Category as classification variable…… | 78 |
| Table 16 | Descriptive statistics of outcome and explanatory variables…… | 81 |
| Table 17 | Pearson Correlation Coefficients of Wildfire_Frequency with selected explanatory variables…… | 82 |
| Table 18 | Multilinear regression results of Wildfire_Frequency with explanatory variables.…….. | 85 |
| Table 19 | Pearson Correlation Coefficients of Incident_Acres_Burned with selected explanatory variables…… | 86 |
| Table 20 | Multilinear regression results of Incident_Acres_Burned with explanatory variables…….. | 89 |
| Table 21 | Analysis of variance for variable Wildfire_Frequency ……..….. | 90 |
| Table 22 | Analysis of variance for variable Incident_Acres_Burned……... | 91 |
| Table 23 | MANOVA output tables …….. | 92 |

Table 24 Supernova results for Regression model example ……. 93
Table 25 Results of XGBoost models for Wildfire frequency and severity …….. 103

## List of Illustrations

Figure 1 Historical fire size and frequency graphs ……. 1
Figure 2 Maximum 1 h (a) and 8 h (b) averaged surface O3 data, and the number of days in exceedance of the state 1 h (c) and 8 h (d) O3 standards, for selected air basins in California ……... 5
Figure 3 Average above-average temperature and precipitation in California …… 6
Figure 4 Population of five largest counties in Southern California ……... 7
Figure 5 Components of research question ……... 8
Figure 6 Map of drought severity in the U.S. …… 10
Figure 7 The Western U.S. forest fire activities and aridity using VPD (hPa) …… 19
Figure 8 Interactive Map depicting Fire Tracking and Air Quality for the State of California …….. 21
Figure 9 Carbon Emission Data depicting changes in different Sectors of California over the years …… 22
Figure 10 Overview of California Carbon Emission for 2021 in comparison to 2020 ….. 23
Figure 11 Current U.S. Drought Monitor Conditions for California …… 25
Figure 12 The North Pacific Jetstream winds ……. 26
Figure 13 Average number of occurrences of Santa Ana events per month ……. 27
Figure 14 Urban developed communities in each county classified as near or within the WUI in Southern California …….. 30
Figure 15 Data mart design ……... 46
Figure 16 The Number of Wildfires Recorded in Southern California …….. 53
Figure 17 Number of Acres Burned in Southern California Due to Wildfires …….. 53
Figure 18 Acreage Burned by Individual Wildfires in Southern California …….. 54

Figure 19 Number of Wildfires Recorded per Month in Southern California Between Years 2013 to 2022 ..... 55

Figure 20 The Total Acreage Burned per Month in Southern California Between Years 2013-2022 ..... 56

Figure 21 Total Number of Wildfires Recorded in Southern California Counties Between Years 2013 and 2022 ..... 57

Figure 22 Total Acreage Burned in Southern California Counties Between the Years 2013 and 2022 ..... 57

Figure 23 Distribution of Metric_Tonnes_of_CO2_Emissions. ..... 62

Figure 24 Distribution of Total_Fuel_Consumption_(MMBtu). ..... 62

Figure 25 Distribution of Metric_Tonnes_of_CO2_Emissions using Sector_Group as classification variable. ..... 63

Figure 26 Distribution of Metric_Tonnes_of_CO2_Emissions using Sector_Group as classification variable..... 64

Figure 27 Distribution of Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable. ..... 64

Figure 28 Distribution of Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable..... 65

Figure 29 Precipitation histogram with descriptive statistics..... 66

Figure 30 Box plot showing precipitation amount by County ..... 67

Figure 31 Histogram of Maximum Temperature ..... 68

Figure 32 Bar chart showing temperature maximum by County ..... 69

Figure 33 Histogram of Minimum Temperature ..... 70

Figure 34 Boxplot showing temperature minimum by County ..... 71

Figure 35 Distribution of TotalPop ..... 72

Figure 36 Distribution of Income..... 72

Figure 37 Box plot of variables TotalPop and County..... 75

Figure 38 Box plot of variables Income and County..... 75

Figure 39 Distribution of Area_SquareKm…….……..…….…….. 76

Figure 40 Distribution of Housing_number…….……..…….……. 77

Figure 41 Distribution of Population_count…….……..…….……..… 77

Figure 42 Correlation matrix of all continuous variables…….……..…….. 80

Figure 43 Scatter plots of outcome variable Wildfire_Frequency with all explanatory variables…….……..…….……..…….……. 83

Figure 44 Scatter plots of outcome variable Incident_Acres_Burned with all explanatory variables…….……..…….……..…….……. 87

Figure 45 Multiple Linear Regression Model of INCIDENT_ACRES_BURNED…… 94

Figure 46 Multiple Linear Regression Model of WILDFIRE_FREQUENCY……... 95

Figure 47 Multiple Logistic Regression Model of INCIDENT_ACRES_BURNED…… 96

Figure 48 Multiple Logistic Regression Model of WILDFIRE_FREQUENCY…… 97

Figure 49 Heatmap of all the continuous variables…….……..…….…… 98

Figure 50 Wildfire Frequency of Test and Predicted Data Comparison…….…… 99

Figure 51 Correlation Heatmap of INCIDENT_ACRES_BURNED before model optimization…….……..…….……..…….……... 100

Figure 52 Correlation Heatmap of INCIDENT_ACRES_BURNED after model optimization…….……..…….……..…….……... 101

Figure 53 Wildfire severity of Test and Predicted Data Comparison…….…… 102

## Chapter 1 - Introduction

### Background

The research conducted on large California wildfires by Keeley et al., 2020 shows that the western region of the United States has a history of large wildfires, and Southern California's recent wildfire conditions are not particularly uncommon. However, the same research also shows that the frequency and severity has increased in pace. Causes for these fires are complex and numerous, but climate change has been implicated as a critical factor. There is historical evidence drought has often been the catalyst of large fires and California has experienced an unprecedented drought in the last decade (Keeley & Syphard, 2021). Figure 1(a) depicts the size of large fires from 1860 to 2020 and Figure 1(b) depicts the frequency of large fires over the same time period.

**Figure 1**

*Historical fire size and frequency graphs.*

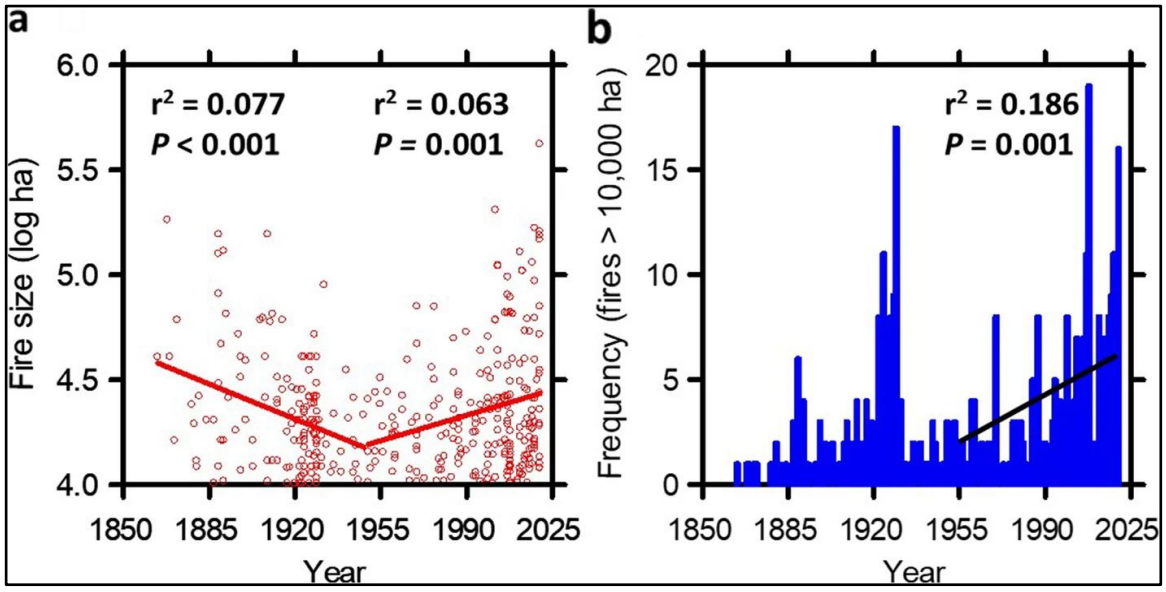



Figure 1 forecasted that incidents of fires will increase in the future as the regression model shows statistical significance based on the $r^2$ and p value. According to the Center for

Climate Change (2022), climate change is often identified as the culprit to increasing wildfires, however, there are many factors that contribute to climate change. The website, hosting the graphic presented in Figure 1, also states that just as there are many reasons for climate change, there are many reasons other than climate change such as populations, carbon emissions, and weather, causing the increase in frequency and severity of the Southern California wildfires.

**Problem Statement**

Researching why the wildfires in Southern California were increasing in frequency and severity was an important step in the fight of protecting the balanced ecosystem in the area and minimizing human and environmental impact (Swain, 2021). According to the world population review, Southern California makes up the southernmost counties in the state of California. The region has a population that is estimated to be around 23,762,904 in 2022 (U.S. Census Bureau, 2022). Most of the region's population resides within the Greater Los Angeles metropolitan area, which has five counties including Los Angeles, Ventura, and Orange with over 17.5 million people living in the region (U.S. Census Bureau, 2022). It is the second-largest combined statistical area in the country, falling only behind the New York metropolitan area (U.S. Census Bureau, 2022). Approximately 60% of the state's total population lives in Southern California (U.S. Census Bureau, 2022).

Not only is the population in Southern California exceptionally large compared to most of the U.S. counties, but the agriculture produced in the area is extensive. Globally crop production needed to increase by the middle of the century to meet the predicted demand for food from increasing population and lifestyle changes. Most of the common crops in California need to incubate for several years before they can be harvested and sold. The prices with which the crops are sold, may vary depending on a few variables such as size, color, chemical

composition, firmness, and aesthetic features. Most, if not all, of these attributes, can vary largely with relatively small temperature changes during the critical development stages of the crops. An evaluation of climate change impacts 8 out of the 20 major permanent crops grown in California showing that temperature variations of 2 degrees Celsius were most closely related to yield reductions in almonds, wine grapes, strawberries, hay, walnuts, table grapes, freestone peaches, and cherries (Pathak et al., 2018).

The population of Southern California are dealing with the aftermath with increasing size, severity, and frequency of the wildfires. Pathak et al. (2018) pointed out that California is the largest and possibly one of the most complicated agricultural states in the country, because of 77,500 farms comprising 5.7 million hectares of pasture and rangeland along with 3.8 million hectares of irrigated cropland. The state plays a critical role in producing approximately a third of the entire country's vegetables and two-thirds of fruits and nuts with a production value of $50.5 billion, this too on just 1.2% of the country's agricultural land. California also presents itself as a production superpower state, in that it produces almost 400 different commodities which cannot be made anywhere else in the nation. Pathak et al. (2018) further emphasized how half of the country's nuts and fruits consumption is grown in-state, comprising items such as dates, figs, olives, kiwis, almonds, and pistachios, something it is highly known for. The state is also a leader in the production of avocados, lemons, plums, strawberries, and more, and in 2015, it's top 20 crop and livestock commodities presented with a monumental gross revenue of more than $41.1 billion (Pathak et al., 2018).

According to Pathak et al. (2018), the economy of Southern California is linked to agriculture, and the wildfires in the region are placing high stress on the agricultural sector. By increasing the understanding of why the wildfires are increasing in size, severity, and frequency

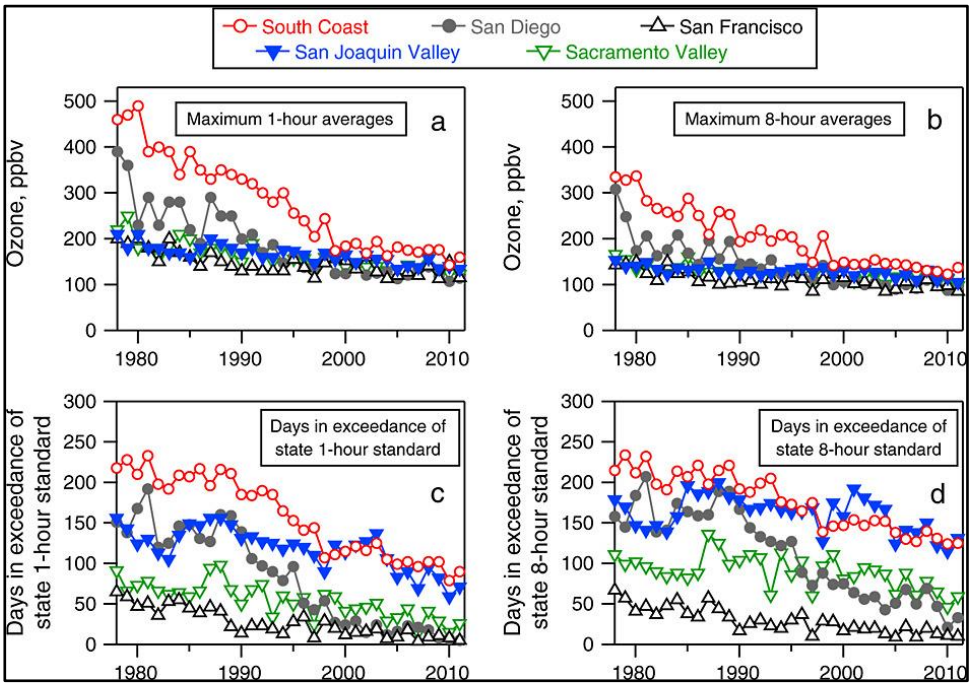
Southern California can make informed decisions as to how to implement remedial change. It is important to know what changes to implement to maximize positive results while minimizing negative side effects. These changes should focus on decreasing those favorable conditions for wildfires. This can be done by identifying what is creating favorable conditions. Only by making informed decisions can the wildfire epidemic be stemmed. Weather, population, and carbon emissions impact wildfire frequency and severity (Swain, 2021; Williams et al., 2019; Pathak et al., 2018). The essential question is, how do these factors affect the impact on people for wildfires, and what can be done to limit their consequences?

**Research Hypotheses**

The effort of this research is to identify how carbon emission, weather, and population impact the frequency and severity of wildfires in Southern California. Carbon emissions are a known source of climate change (Center for Climate Change 2022). Over the past several decades in the U.S., emissions reductions implemented for vehicles and point sources have significantly improved air quality in most metropolitan areas. Since 2013, the rate of improvement in air quality in most regions of the U.S. had slowed, both in terms of regional ozone concentrations and ozone exceedance days e.g., Figure 2 for California (Ryerson et al., 2013).

**Figure 2**

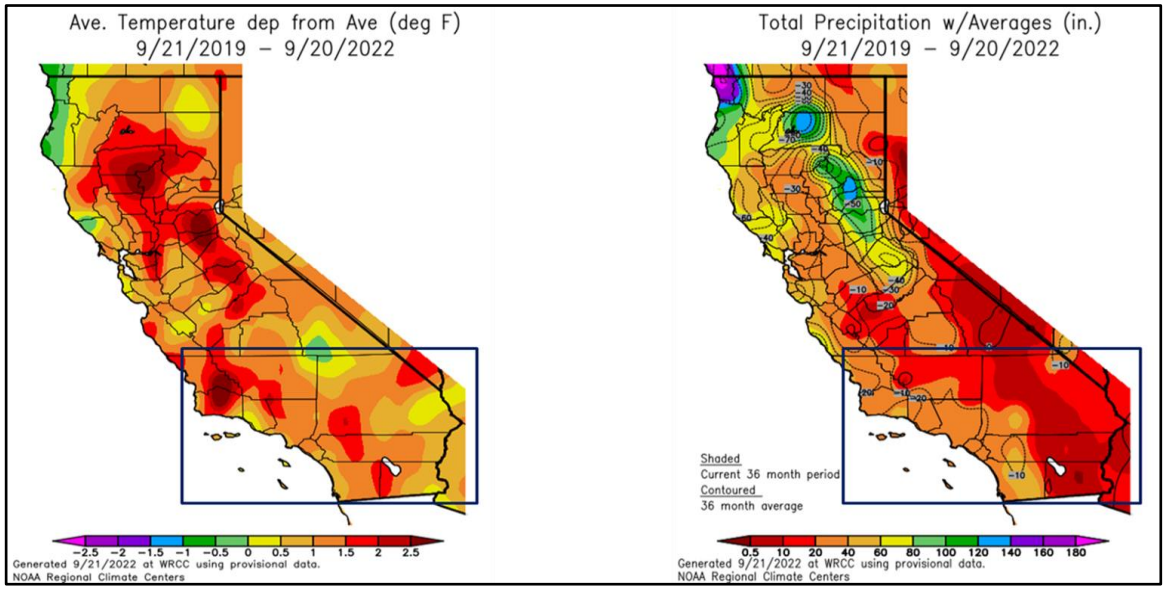
*Maximum 1 h (a) and 8 h (b) averaged surface O3 data, and the number of days in exceedance of the state 1 h (c) and 8 h (d) O3 standards, for selected air basins in California.*




Weather is also a projected cause of wildfires. Weather is responsible for the dry conditions of Southern California. Moisture in California is largely controlled and moved by the strength and position of the North Pacific Jet (NPJ) stream, high-altitude winds that blow into the state from the west during the cooler rainy season. The strength and position of the winds influence regional conditions that carry over into the warmer dry season when wildfires are more prone to occur. The wet-season NPJ thus becomes an important precursor of summer fire conditions (NOAA. 2019). The average temperature and precipitation levels have therefore been hampered because of an inefficient NPJ. Causing severe dry conditions as depicted in Figure 3.

**Figure 3**

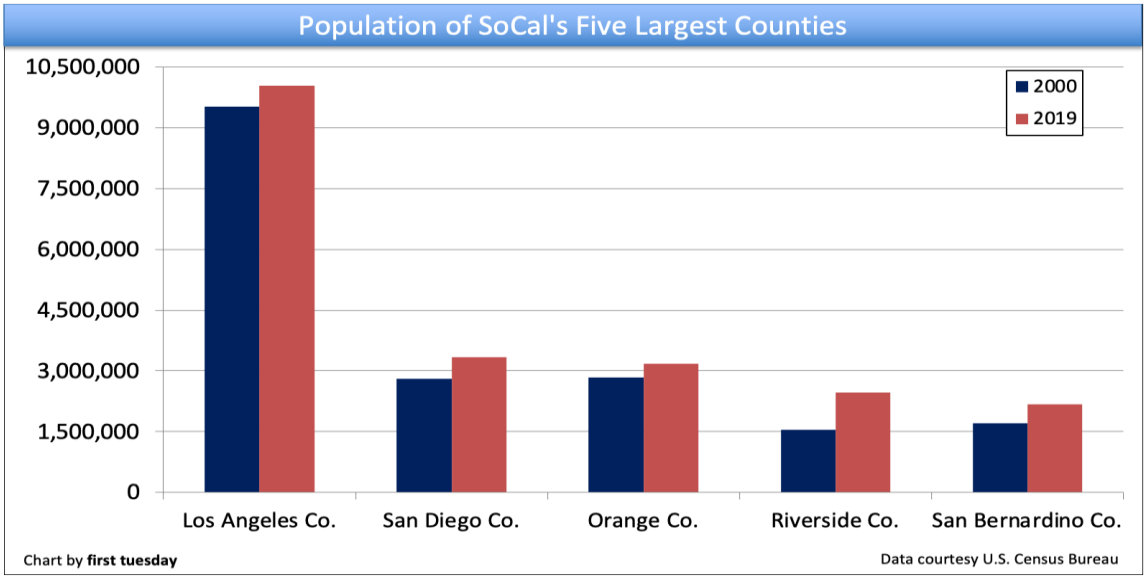
*Average above-average temperature and precipitation in California*




The influence of population cannot be overlooked when discussing the wildfires in Southern California. With such a dense population the urbanization bordering the fire favorable terrain makes accidental fires a real concern. Figure 4 is taken from Firsttuesday Journal that shows population growth in Los Angeles, San Diego, Orange, Riverside, and San Bernardino counties of Southern California between the years 2000 and 2019.

**Figure 4**

*Population of five largest counties in Southern California*



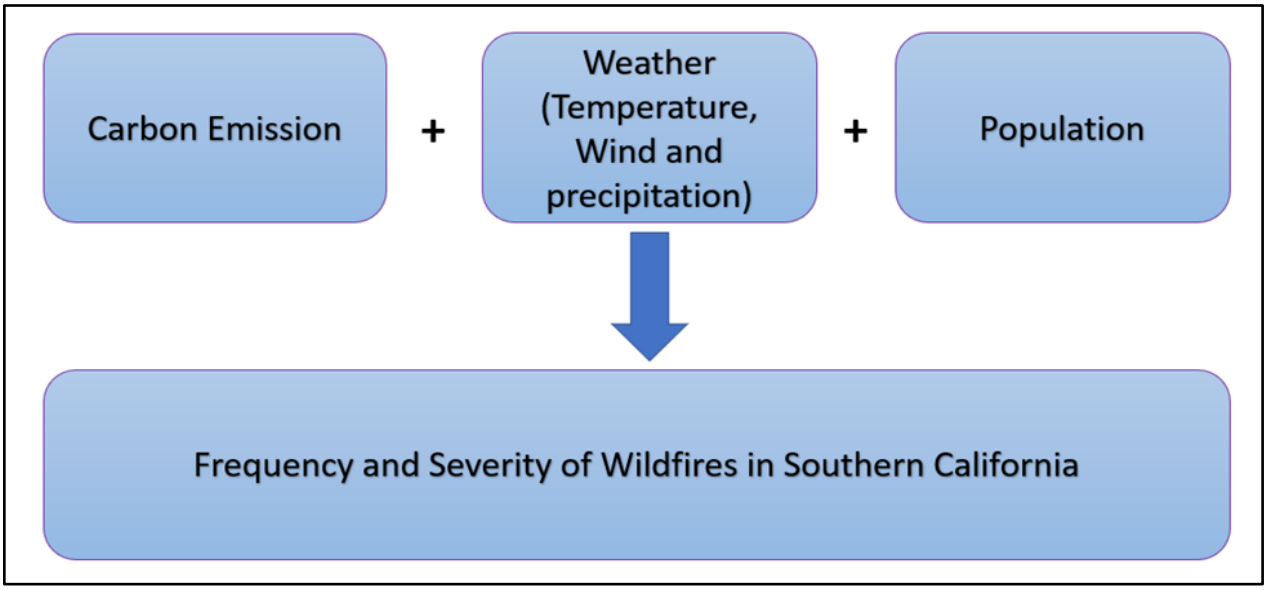

By testing how carbon emissions, weather, and population affect the wildfires of Southern California. This research can identify or rule out primary factors in the wildfire frequency and severity in Southern California.

**Objectives**

Increases in the number of wildfires in Southern California are affecting human health, infrastructure, and ecosystem management to a large extent. Despite the consideration of meteorology and wildland fuels as important controlling factors for wildfires worldwide, there is still a dispute over their significance (Jin et.al., 2015). Figure 5 has been created by A-Team to represent how the research question has been divided into four components: carbon emission, weather, and the population established as explanatory variables in this study, and frequency and severity of wildfires are outcome variables.

**Figure 5**

*Components of research question*




***Carbon Emission***

Carbon emissions are usually classified as chemical emissions originating from the burning of fossil fuels, whether in a solid, liquid, or gaseous state. These emissions are also said to be one of the biggest contributing components of climate change. According to the California Air Resources Board, in an effort to further understand these emissions in regard to wildfires, the board has been tracking and finally released an estimate of wildfire carbon emissions ranging for the 2000-2019 period. California in the past couple of decades has been subjected to rapidly deteriorating climate conditions, and with the high frequency and magnitude of wildfires spreading across the entire state, the situation has become even more significant to be able to predict and prepare for whenever the next wildfire is going to occur. Carbon emission is supposed to be one of these predictor variables. Abnett (2021) states around 1.76 billion tons of carbon emissions exist globally for the year 2021. Moreover, California experienced its largest recorded fire known as Dixie the same year. Alarmingly, looking at North America, fires in

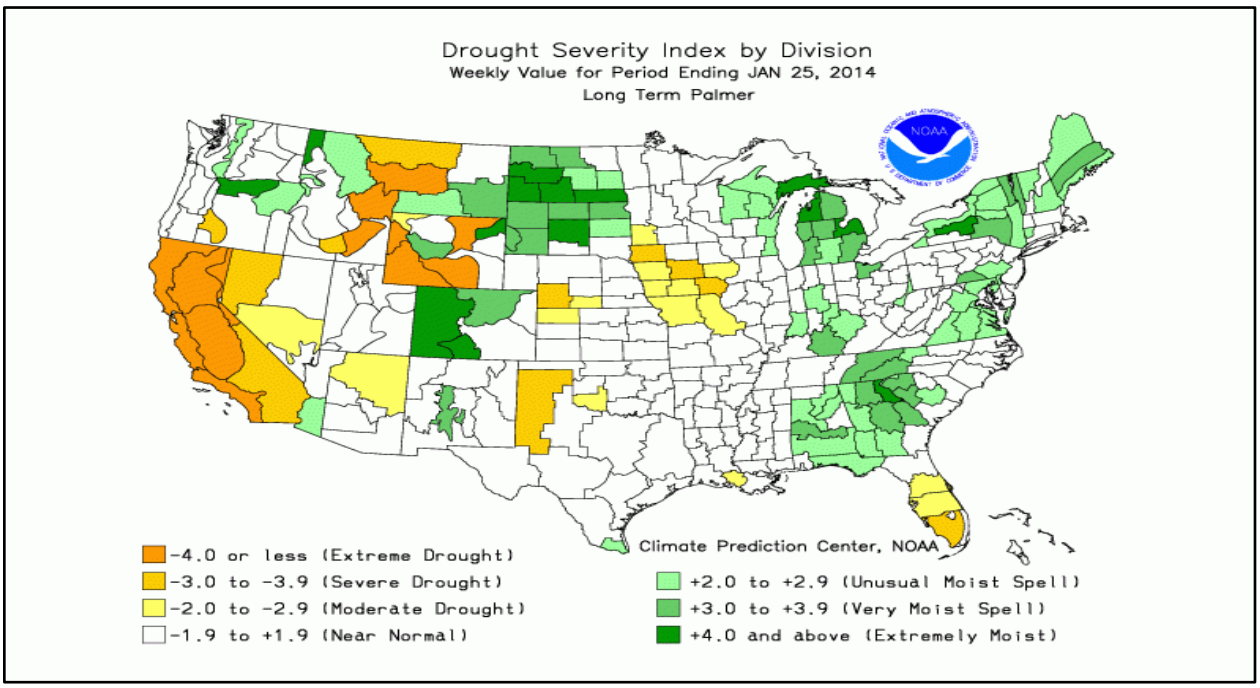
Canada, California, and the U.S. Pacific Northwest were said to have emitted around 83 million tons of carbon, releasing huge smoke plumes which drifted across the Atlantic all the way to Europe (Copernicus, as cited in Abnett, 2021). Being able to predict wildfires with the amount of carbon emission present in an area is becoming highly necessary.

***Weather***

The weather in Southern California is commonly thought of as dry and arid. Although, the level of dryness in recent years is above average. However, very recently, 2017 showed a change in the pattern seen in the longer record. The severe Tubbs and Thomas fires of 2017 were a high-precipitation year. This year overrode the NPJ's historical relationship with low-fire extremes after cool seasons of very high moisture. Extreme precipitation had compromised the Oroville Spillway earlier that year in addition to bringing about dangerous floods and landslides. Prior to modern fire suppression, the paleoclimatic reconstruction showed no cases of a high-precipitation year coupled with a high-fire year. If this trend in warming continues, as is the scientific consensus, then any significant wet season rain and snow may not ensure a quiet fire season afterward. Besides fire risk and its associated health and economic impacts, such a change could alter species distribution, forest composition, and ecosystems (NOAA. 2019). If the NPJ is no longer providing adequate wet conditions conducive to rainfall in the Southern California region then the area would be more prone to incidents of Wildfires. If this is the case, then it is possible that this is a larger historical climatic cycle and is not out of the normal. It then becomes important to understand what can be done. Figure 6 depicts NOAA's drought severity index as of 2014, which displays that most of California, including Southern California, is in a state of extreme drought.

**Figure 6**

*Map of drought severity in the U.S.*




Depending on whether the wildfires in Southern California are becoming more frequent and severe based on weather is important to understand. The root problem of any issue must first be understood before safeguards or solutions can be formulated. Weather is likely the most difficult variable to try and predict and control regarding wildfires. but as stated before, knowing is the first step.

***Population***

Over 17.5 million people live in Southern California that has extensive undeveloped land, and four National Forests along with populous cities such as Los Angeles and San Diego included in the region (Jin, et al.,2015). The migration of the population to the suburbs is the source of an uncontrolled expansion of urban areas that escalated the number of human structures into wildlands, forests, and habitats. Natural ecosystems are impacted by human

activities to a great extent with the increasing establishment of residential communities on the outskirts of a city. Wildland-Urban-Interface (WUI) areas are defined to observe and assess if human activities have any impact on the local climate and environment. Because of the accumulation of wildland vegetation and concentration of flammable human structures along with sparks scattered in a dry area by human activities, WUI in California is recognized as a high-risk area for human-caused wildfires (Li et al., 2022). The possibility of humans sparking fire is increasing with disproportionate population growth in the areas with hazardous vegetation and the risk may be higher than what was previously understood (Rao et al., 2022). The housing wealth and other economic indicators show a prominent imbalance between the households in or adjacent to fire-prone zone areas and those near the coast (Jin et al., 2015). Receiving any information well in advance about potential wildfires can be of great help to develop an effective mitigation plan for wildfires. The research to predict wildfire frequency and severity will create an opportunity to improve an understanding of regional fire dynamics in Southern California. The objectives of this research are to understand the pattern of population growth in wildland-urban-interface areas, assess if variation in household income has any influence on wildfire mitigation, and identify if population growth in WUI does have any association with the frequency and severity of wildfires in Southern California.

## Limitations of the Study

### *Carbon Emission*

Carbon emissions from the use of fossil fuels across several industries have always been a huge cause of climate change. Moreover, climate change itself has been known to be a key factor in increasing the risk and extent of wildfires in the Western United States (Center for Climate Change 2022). According to the California Air Resources Board, wildfires themselves tend to also contribute to climate change by releasing carbon dioxide (CO2) emissions and other greenhouse gasses (GHG). This presents a never-ending cycle, contributing to more incidents of wildfires relevant to carbon emissions. However, even as implied it might be, an explicit inter-dependency cycle has not been stated as part of this study.

### *Weather*

Much of the weather data, retrieved during a search for sources, is broad in focus and rarely specifically on the Southern California area. It is often times that a broader study was conducted on the state of California. This means concepts can become extrapolated and used to apply only to the Southern California region. Furthermore, local weather changes frequently based on year and location. There may be data that exist showing nothing of concern. One example is from the article, Revisiting the recent California drought as an extreme value, stating, spatially weighted averages of the Palmer Drought Severity Index (PDSI) over central and Southern California show that the 1-year 2014 drought was not as severe as previous droughts (Robeson, 2015). While some of the data in the above-mentioned study show less severe drought conditions, when observed as a whole it is reinterpreted to be an anomaly in the record, but still exhibited drought conditions.

### *Population*

According to Coughlan et al., 2019, social and ethnic backgrounds of people, cultural and psychological relationships to fire and land management, social capital, and level of trust in government are considered as diversity factors related to populations. These diversity factors were not included in this study to determine if population growth and their income have any impact on wildfires in Southern California. Including "ethnic backgrounds and cultural or psychological relationship with fire" may add more depth to the study, however, it may produce results that are biased toward certain groups of people, such as people living within rural habitats.

### *Wildfire*

When researching wildfires, the scope of the research was limited to "frequency" and "severity" of the wildfires, where severity described the spread of the wildfires in acreage burned and frequency described the number of occurrences over specific periods of time. Due to this approach, several limitations are acknowledged such as vegetation type and data quality. For the purpose of this study, vegetation type was defined as the specific type of plants that grow in the area. Vegetation type was found to have a significant impact on wildfire intensity after ignition, as different types of plants such as black sage, chamise, greasewood, coyote brush, and others native to California are found to increase flammability and the propensity of fires to spread (Fire Safe Marin, 2021).

Data quality refers to the data granularity of the weather and fire data collected. The data used in this research is on the county level. For future research activities, to improve wildfire prediction accuracy, our team recommends using zip code level data or data collected by satellite imaging with 27 - 28 km resolution (Shmuel & Heifetz, 2022). Due to these limitations, future

research could further explore the impact of wildfires by considering additional factors such as the one mentioned.

## Definition of Terms

**Table 1**

*Definition of Terms*

| Terms | Definitions |
|---|---|
| **AFAB** | Annual Forest Area Burned: The total amount of acreage burned annually (Juang et al., 2022) |
| **KBDI** | Keetch-Byram Drought Index: Describes the upper 8 inch soil moisture level accounting for daily maximum temperature, water evaporation rate, the number of days since last precipitation and the amount of the precipitation (Brown, 2021) |
| **Meteorology** | The climate and weather of a region (Glickman, 2000) |
| **MMBtu** | One million British Thermal Units, a thermal unit of measurement for Natural Gas (U.S. Energy Information, 2022) |
| **MTCO2** | Metric tons of carbon dioxide, a metric measure used to compare the emissions from different greenhouse gasses based upon their global warming potential (U.S. Energy Information, 2022) |
| **NERC Region** | North American Electric Reliability Corporation, not-for-profit international regulatory authority whose mission is to assure the effective and efficient reduction of risks to the reliability and security of the grid (U.S. Energy Information, 2022) |
| **NFIRS** | National Fire Incident Reporting System U.S. Fire Administrator. (2022) |
| **NPJ** | North Pacific Jet Stream: A wind flowing off the pacific ocean that can control humidity levels on the West coast of the United States (National Oceanic and Atmospheric Administration, n.d.) |
| **PDSI** | Palmer Drought Severity Index: Uses readily available temperature and precipitation data to estimate relative dryness (National Oceanic and Atmospheric Administration, n.d.) |
| **SPI** | Standard Precipitation Index: widely used index to characterize meteorological drought on a range of timescales (National Oceanic and Atmospheric Administration, n.d.) |

| | |
|---|---|
| **UTC** | Coordinated Universal Time: also known as “Z time” or “Zulu Time”: It is set to mean solar time at $0^0$ longitude and is not being adjusted for daylight saving time (WFIGS - Wildland Fire Locations Full History, n.d.) |
| **VPD** | Vapor Pressure Deficit: The amount of moisture difference in the air between the observed and the maximum amount of moisture the air can hold, saturated air (Williams et al., 2019) |
| **Wildland Fuel** | All kinds of plant material that can act as fuel during wildfire, including grasses, shrubs, trees, dead leaves, and fallen pine needles (Office of Wildland Fire, n.d.) |
| **WRD** | Wetting Rain Days: Days with precipitation greater than 2.54 mm (Williams et al., 2019) |
| **WUI** | Wildland-Urban-Interface: The WUI is the zone of transition between unoccupied land and human development. (U.S. Fire Administration, n.d.) |

**Summary**

Sources of California’s Wildfires consist of different and connected parts; climate change is an extremely important factor of all. Having said that, changing climate is not the only responsible element for increasing wildfires in Southern California. Carbon emission, Weather, and Population also contribute to the ignition and spread of wildfires. The focus of this study is to summarize the statistics that predict the frequency and severity of wildfires in Southern California as more than half of the population lives in this region of the state. By obtaining information about potential risk of wildfires beforehand, local authorities and households in fire-prone areas can design and implement a mitigation to avoid these wildfires or to reduce losses. Population growth seems to be one of the impacting components of the increased amount of carbon in the air which is a known source of extreme weather conditions. Although the factors of population, carbon emission, and weather are connected to each other in some respect, each concept has been studied to understand the existing circumstances of individual components with respect to wildfire.

# Chapter 2 - Literature Review

## Wildfire Frequency and Severity

Wildfires, as a type of fire, are classified as unplanned and uncontrolled fires affecting wildland vegetation. Wildfires can be exceedingly dangerous and destructive, especially in areas such as California where there is a high risk of expanding burn areas to nearby structures and cities. Between 2012 and 2021, an average of 61,289 wildfires a year were reported, engulfing over 7.4 million acres of land (CRS, 2022). Looking deeper at the data nationwide, 58,698 wildfires were recorded in 2021 alone, resulting in over 7.1 million acres of total damage. Of the 58,698 wildfires, 9,281 were recorded in California, alone, resulting in over 2.23 million acres of damage (Statista, 2022). At 31%, California endured the most wildfires of any state, with Texas as the next closest state logging 5,567 wildfires, and North Carolina with 5,151 (Statista, 2022a).

Looking at the nationwide wildfire statistics for the past 30 years, overall, data shows that the average wildfire frequency has slightly decreased (CRS, 2022). However, further review of the data shows that the average acreage affected during the same period has increased (CRS, 2022). Comparing the frequency of fires to acreage burned California's annual wildfire statistics display a similar pattern for the number of acres burned. The calculated five-year rolling average of acreage burned shows an approximately 3-fold increase of the annual acreage destruction between the years 2011 and 2021 (Fleck, 2022).

As of September 2, 2022, there were over 6,000 incident reports with estimated 350,000 acres affected (2022 Fire Season Outlook, n.d.). The largest wildfire was recorded on July 29th, 2022, engulfing over 60,000 acres of land (Statista, 2022b). In Southern California, Los Angeles County experienced one of the largest wildfires in 2018, affecting 96,949 acres and 31,089 acres in 2020.

The observed increase in wildfire frequency and severity in California, which is highly seasonal, reaching its peak during summer months, has been attributed to the increase in summer atmospheric temperature and decrease in rainfall and winter snowpack (Holden et al., 2018). The increase in warm-summer days leads to aridity and atmospheric vapor pressure deficit (VPD). As Williams et al. (2019) report, the centennial increase in aridity and VPD is consistent with the results of climate model simulators for anthropogenic climate changes and trends. The research studied the relationship between California wildfire activities and a variety of anthropogenic climate variables such as aridity defined by VPD, and daily maximum temperature (Tmax) (Williams et al., 2019). The research reports a high degree of correlation between the wildfires and the warming of summer days through the increase of aridity, especially in northern California regions. The research also reported a weak correlation between aridity and wildfires in the non-forested region (Williams et al., 2019).

Published research suggests an exponential dependency of summer forest wildfires on the increase of VPD, in contrast, for non-forest regions, this relationship was inconclusive (Williams et al., 2019; Juang et al., 2022). However, the studies conducted by Holden et al. (2018) to identify the primary drivers of wildfire acreage burned in the Western United States, showed that the number of wetting rain days (WRD) had a 2.5 times greater net effect than the VPD, while winter snowpack had substantially lower impact than WRD and VPD. Williams et al. (2019) research also analyzed the warm weather variation during the fall season which suggested that the increase of fuel dryness dictated by the decrease of fall precipitation and increase of atmospheric temperature led to an additional 8 days per month with a high probability of wildfire for October to December (Williams et al., 20219).

Daniel Swain (2021) reports that impact of the climate change led to a dramatic shift in California's precipitation seasonality causing more shorter and sharper rainy seasons which, coupled with the rise of atmospheric temperature, created perfect conditions for vegetation-drying and forming downslope winds during late fall, all of which increase fall wildfire spread potential (Swain, 2021). Thus, a slight increase in the atmospheric temperature of warm-season days due to the greenhouse effect can have a tremendous impact on California's wildfire activity in the forested and wildland regions. The extreme dry vegetation which is a result of decreased direct precipitation or increased VPD along the delayed and sharpened rainy seasonality created a conducive environment for more frequent and more devastating wildfires in California (Williams et al. 2019; Holden et al. 2018; Swain, 2021).

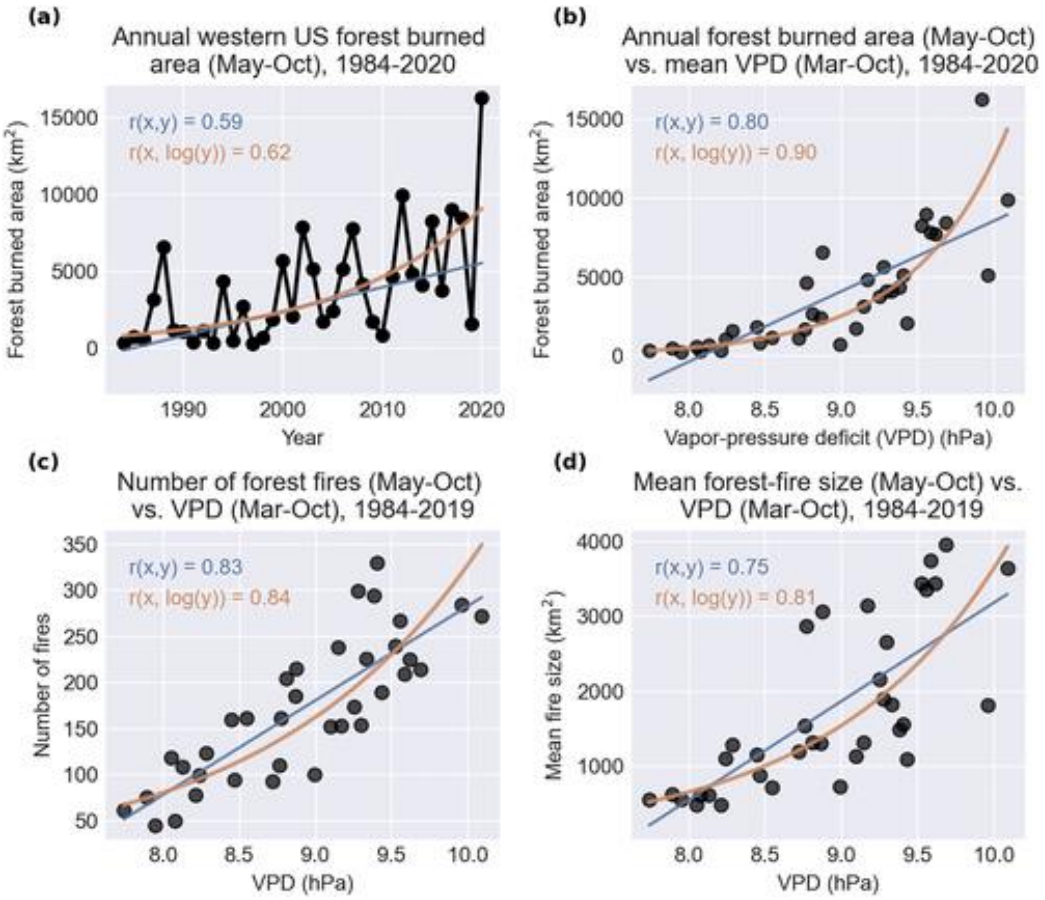
The Juang et al., 2022, research publication studied the positive exponential impact of annual forest area burned (AFAB) in the western U.S. as a function of increased aridity by utilizing a new dataset of daily fire growth data collected by satellite imaging. The results indicated that the wildfire frequency and duration had a positive linear response to the increased aridity, whereas the AFAB displayed exponential growth (Juang et al., 2022). The study also points out that aridity has a greater impact on larger fires due to their extensive firelines, thus, exponentially increasing the potential for fire growth (Juang et al., 2022).

The results of the research publication of Juang et al. (2022) show both a linear and log-fit performed on each analysis (Figure 7). The time series for forest acreage burned, shows the trendline fit both linear and logarithmic were similar ($r = 0.59$ vs. 0.62, and $p<10^{-4}$) leading to a 1,574 $Km^2$ increase of western U.S. AFAB per decade (Figure 7(a)) (Juang et al., 2022). Similarly, Juang et al. (2022) reported a linear trend between the number of fires and VPD (hPa)

with r = 0.83, $p<10^{-4}$ (Figure 7(c)). In contrast, Juang et al.'s results (2022) display a strong positive exponential fit between AFAB and VPD (log-fit with r = 0.90, $p < 10^{-4}$).

**Figure 7**

*The Western U.S. forest fire activities and aridity using VPD (hPa).*




*Note.* Figure shows the research results of Juang et al., published in 2022.

Forest services such as Canadian and U.S. forest services including the U.S. Army utilize widely used Keetch-Byram Drought Index (KBDI), and Canadian Fire Weather Index (CFWI), instead of other fire indices such as Armstrong and Baumgartner, to determine the degree of wildfire susceptibility of the area for a given time (Hamadeh et al., 2015, as cited in Brown, 2021). Since there is a nonlinear relationship between the wildfire frequency, the fire indices,

and meteorological variables, the non-linear indices such as KBDI and CFIWI which account for the cumulative impact of multiple variables, offer a more accurate prediction of possible wildfire activities (Hamadeh et al., 2015, as cited in Brown, 2021).

KBDI calculates the degree of the upper soil moisture for up to 8 inches deep, accounting for the cumulative annual precipitation, and evaporation rate to determine the soil as well as vegetation dryness (Brown, 2021). The calculation of the KBDI index also accounts for the number of days since the last precipitation and the amount of precipitation (Brown, 2021). KBDI calculation also assumes a high correlation between high annual precipitation and more vegetation availability as a burning fuel (Brown, 2021).

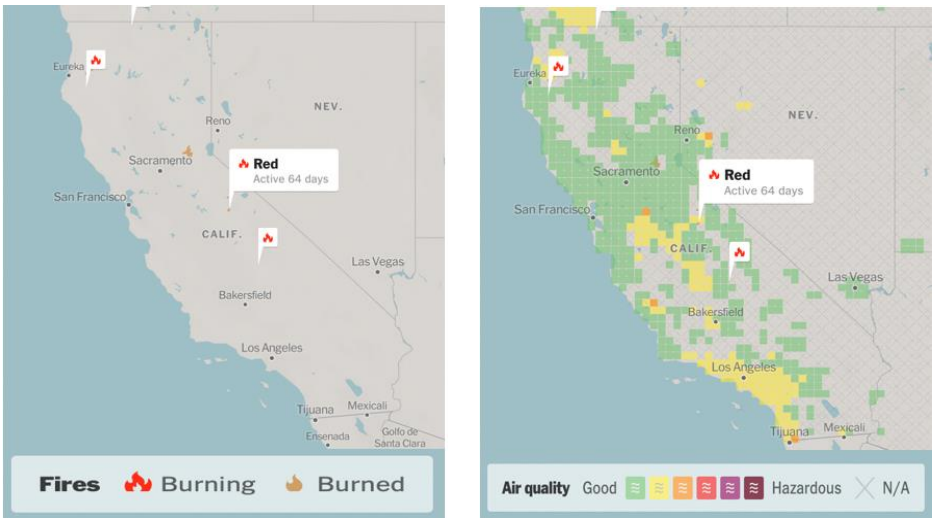
Due to the above-described non-linear interaction between wildfire activities, meteorological variables, fire indices, fuel characteristics, and other variables, Shmuel, and Heifetz (2022) claim that machine learning (ML) produces much more accurate prediction models compared to simpler statistical models such as linear regression. Furthermore, the authors argue that the use of a satellite-based global wildfire dataset with millions of wildfire observations allow more effective and improved ML model development in contrast to the use of a smaller single regional data (Shmuel & Heifetz, 2022). In this research, the multilayer perceptron (MLP) neural network and logistic regression (LR) models were used to test the wildfire prediction accuracy, along with the random forest (RF) and eXtreme Gradient Boosting (XGBoost) ML models as these are ideal for relational data (Shmuel & Heifetz, 2022). The results of this research report 70% wildfire prediction accuracy with the linear regression model, whereas, ML models performed with much higher accuracy, over 90% accuracy (Shmuel & Heifetz, 2022).

## Carbon Emission and Wildfires

While wildfires are not a new topic or phenomenon and their frequency has definitely changed over the last few decades, there are contributing factors - Carbon Emission being one of them. Air quality contributed very much by carbon emissions and wildfires go hand in hand. According to an Interactive Fire and Air Quality Map by the New York Times (2022), the previously burnt wildfire areas tend to have a poor quality of air when compared to areas that have not been burned. For the Bobcat fire near Los Angeles, the wildfire around the end of 2020 threatened parts of nearby foothill communities, where evacuation orders were issued (Bloch et al., 2020). And the  U.S. Forest Service stated that extremely dry brush in areas with "little to no fire history" had fueled the fire (Bloch et al., 2020).

**Figure 8**

*Interactive Map depicting Fire Tracking and Air Quality for the State of California.*



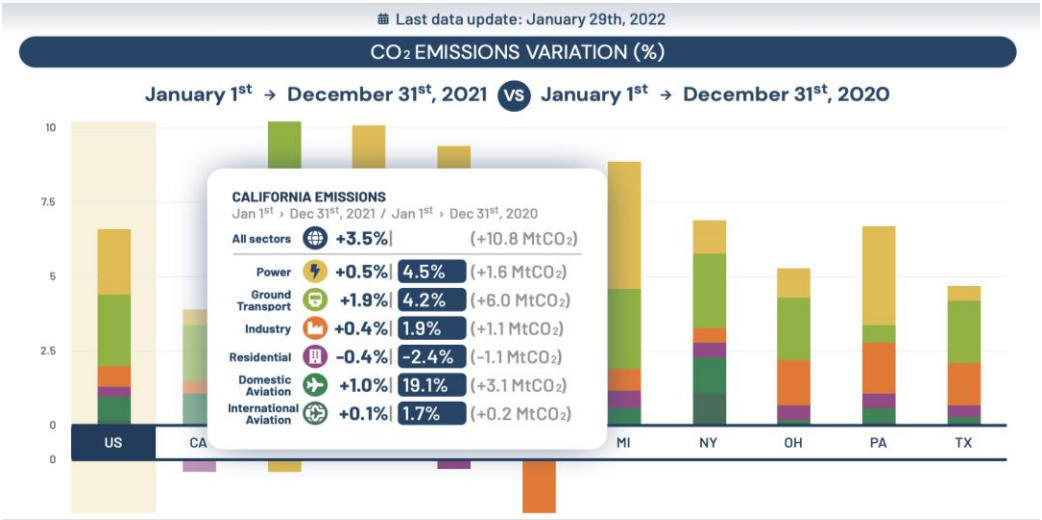

In Figure 8, a viewer can see a correlation of bad air quality after a wildfire has taken place. According to CalFire, the Howard fire incident has been the most recent fire incident located in Ventura County of the Southern California Region as of October 8th, 2022, and it has

affected  the Rose Valley Rd and Gene Marshall-Piedra Blanca Trail, which is located northeast of Ojai area, and its cause has not been determined yet and is currently under investigation.

The U.S. Carbon Monitor states that carbon dioxide emissions from the use of fossil fuels are the primary cause of climate change. This cause is a part of an international initiative that provides regularly updated, science-based estimates of daily CO2 emissions of individual U.S. states. Interestingly, comparing year-to-year data reveals the effects of the COVID pandemic and patterns of sectoral emissions during the recovery of carbon emissions. Figure 9 focuses just on California as a state, domestic aviation and power are the two highest sectors to have increased their carbon emissions when compared with last year, respectively at 19.1% and 4.5% sector growth.

**Figure 9**

*Carbon Emission Data depicting changes in different Sectors of California over the years*

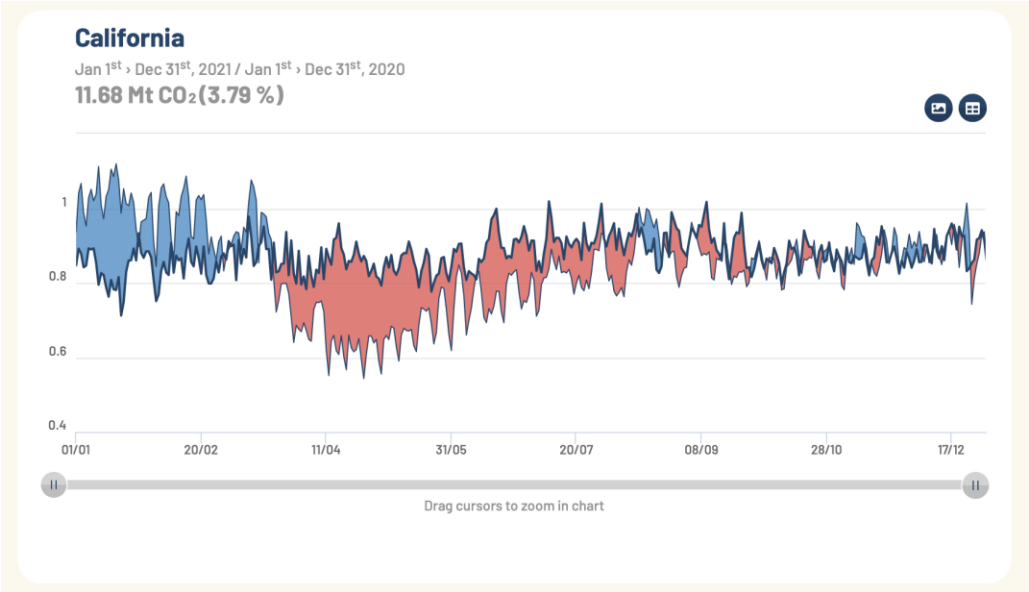



California, overall as a state, has been adopting several changes to make a positive impact on carbon emissions recently, with keeping climate change and wildfires in mind. U.S. Carbon Monitor states that California has decreased its overall carbon emission to a lower capacity than

that of 2020. Figure 10 presents that most of the year's duration has a much smaller amount of carbon emission issued.

**Figure 10**

*Overview of California Carbon Emission for 2021 in comparison to 2020*




*Note.* Blue area map shows decrease, and Red shows increase in carbon emission

The U.S. Energy Information Administration provides us with emission data by individual plants and regions. The data is centered around CO2, SO2, and NOx emissions, for the years 2013 to 2020. It also keeps track of the Sector Code, Prime Mover, Fuel Code, Aggregated Fuel Group, Generation (kWh), Useful Thermal Output (MMBtu), Total Fuel Consumption (MMBtu), Fuel Consumption for Electric Generation (MMBtu), Fuel Consumption for Useful Thermal Output (MMBtu), Quantity of Fuel Consumption, Fuel Units, Metric Tons of CO2 Emissions, NERC Region, Balancing Authority Code, Balancing Authority Name and EIA Balancing Authority Figures, etc.

According to Roper (2020), and the data collected by InsideClimate News, there are five

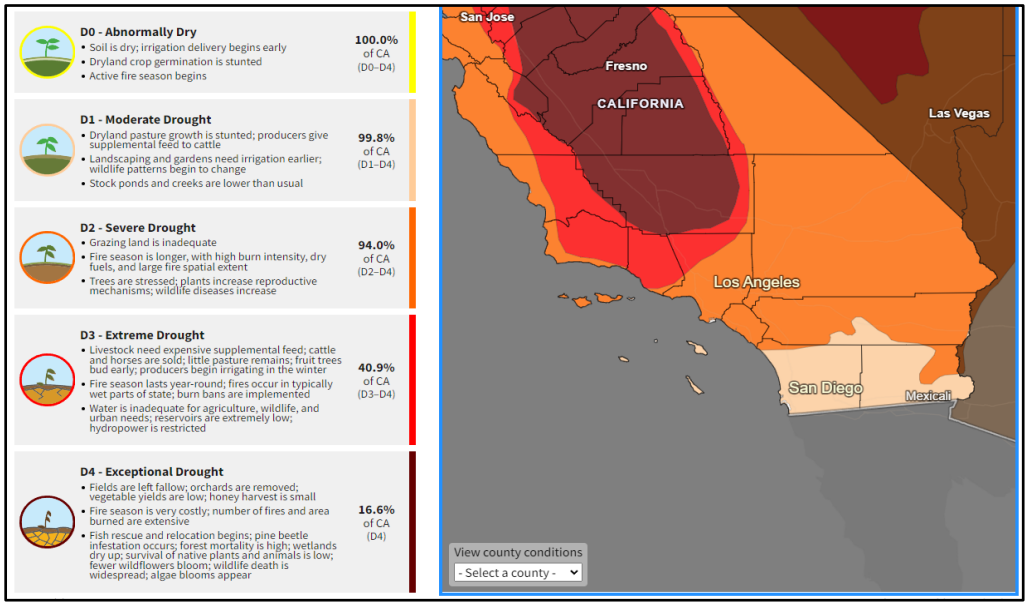
states in the U.S. that have passed legislation for transitioning towards a renewable/carbon-free infrastructure in the coming decades. These states include California, Hawaii, and Washington, which have passed legislation designed to target 2045 as the year each state will be carbon-free. Other states have similar target years as of 2040, 2045, and 2050, while Washington D.C. has the target year of 2032 to be using 100 percent renewable energy. Puerto Rico has also passed legislation requiring the territory to pursue clean energy, while Maine and Nevada have enacted executive orders that set goals for each state to be 100 percent renewable or carbon-free by 2050.

**Effects of Weather on Southern California Wildfires**

The weather's effect on wildfires is a common topic. There have been several published scientific articles based on similar premises. Most of these research articles focus on drought conditions and lack of rain. Interannual variations and the persistence of drought conditions are profoundly important for human activities and natural ecosystems. Recent drought conditions in California have been particularly intense and have been analyzed using instrumental data, paleoclimatic proxies, and modeling approaches (Robeson, 2015). The weighted averages of the Palmer Drought Severity Index (PDSI) are often used when measuring drought levels. Other types of measurements include U.S. Drought Monitor (USDM) as seen in Figure 11, which is updated on a weekly basis.

**Figure 11**

*Current U.S. Drought Monitor Conditions for California.*



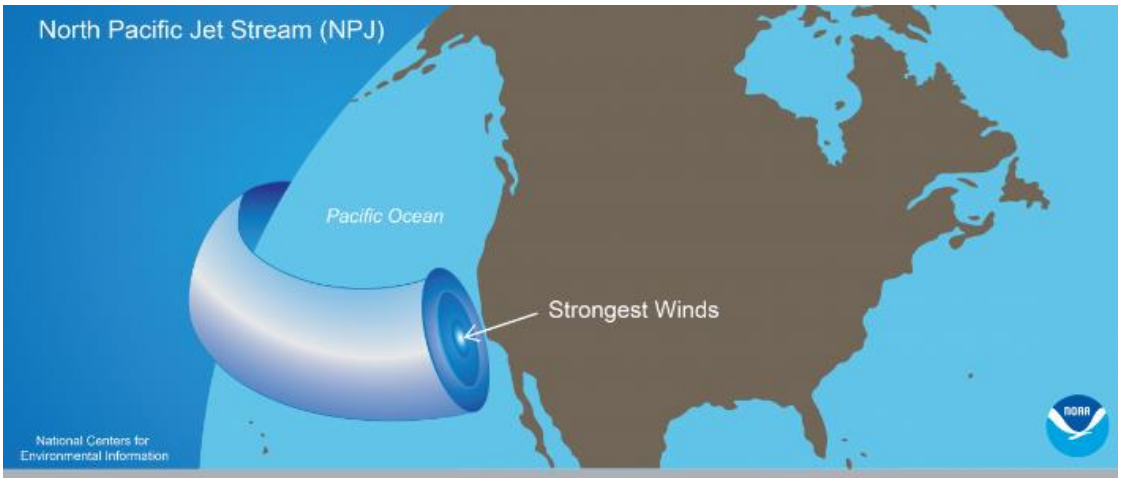

These studies and monitoring agencies all seem to indicate wildfires are linked to drought conditions. The great majority of large fires in Southern California occur in the autumn under the influence of Santa Ana windstorms. These fires also cost the most to contain and cause the most damage to life and property. Spring and summer fires in Southern California regions are usually easier to contain because of higher fuel moisture and the general lack of high winds. However, some fires burn in a remote wilderness area of rugged terrain that makes access difficult. Coupled with this was severe drought generating fuel moisture levels considerably below normal for early summer (Keeley et al., 2009).

Drought is not the only weather factor that could play a role in the wildfires in Southern California. This is because conditions fostering large fall and winter wildfires in California are the result of large-scale patterns in atmospheric circulation, the same dangerous conditions are

likely to occur over a wide area at the same time (Westerling et al., 2004). The North Pacific Jet Stream (NPJ) is responsible for the delivery of moisture to the west coast of the U.S.

**Figure 12**

*The North Pacific Jetstream winds.*



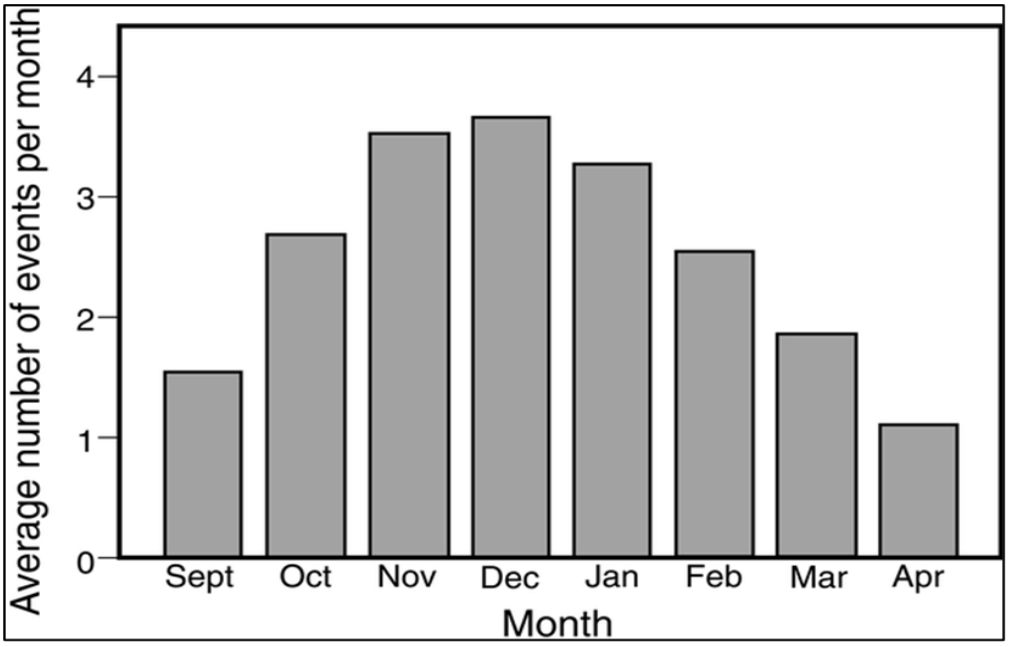

The North Pacific Jet (NPJ) travels eastward at variable wind speeds and directions toward California at an altitude of about 11 kilometers above the ocean's surface. The strength and position of the winds take on importance in relation to the amount and intensity of moisture the jet stream delivers. This graphic represents a winter-average path of entry to California that could produce a very-wet, low-fire season in the state. Courtesy of NOAA NCEI.

Atmospheric conditions create a unique and difficult object to track and predict. There are occurrences of events described as Santa Ana winds. The seasonal Santa Ana wind is a hot, dry, foehn-type, easterly, or northeasterly wind that blows from the deserts east of the Sierra Nevada to the coast of Southern California (Glickman, 2000). It tends to occur in winter and spring. While it is named after the pass and river valley of Santa Ana, California, it can affect much of the Southern California region. The occurrence of the Santa Ana wind is anticipated each year. It is an important local, meteorological phenomenon commanding scientific (e.g., Mensing et al.,1999) and social (e.g., Miller, 1968) interest because of its relationship to risk of forest fires (e.g., Minnich, 1983; Keeley and Fotheringham, 2001), and therefore to watershed

runoff, its effect on temperature, humidity, and on the distribution and deposition of air pollutants.

**Figure 13**

*Average number of occurrences of Santa Ana events per month*




While NOAA is the largest organization currently studying weather's effects on wildfires, they are by no means the only organization. The American Meteorological Society, American Geospatial Union, National Center for Atmospheric Research, U.S. Forestry Services, Cambridge University, and the Society for Conservation Biology to name a few. These organizations' and universities' research proves that weather, in one aspect or another, is linked to the wildfires in Southern California. The research conducted by these institutions typically spans centuries using historic data to quantify the level of drought in the Southern California region resides.

The gap in the research released by these institutions lies in whether the wildfire severity and frequencies are directly related to the weather conditions in Southern California. While

predictive analysis on weather will be difficult to achieve, it is possible. Meteorologists conduct forecasting regularly to warn people of impending storms and other phenomena. The uniqueness of this research project will be to identify if wildfires are indeed a variable in Southern California wildfire frequency and severity, then to attempt to identify predictors to aid in future forecasting of large-scale wildfires in Southern California. If achieved, firefighting teams and other first responders can have larger warning times to prepare and pre-stage equipment. This will allow them to prepare the area with fire retardant before an actual fire is ablaze. Ultimately, it will save lives, time, and money. By keeping the correct weather conditions monitored and predicting possible wildfires this model could be adapted to other parts of the world. With a bit of fine-tuning, it may be used to predict floods instead of wildfires. Aiding in catastrophes like the massive flooding in Pakistan. None of the mitigations happens unless the correct predictive weather models are created and improved.

**Population growth in Wildland-Urban Interface area**

Wildfires are frequently affecting areas beyond the traditional wildlands as residential growth has expanded into these areas (U.S. Fire Administration, 2022). Forests and communities are connected in many ways; parks, trails, rivers, and vistas remind us of the beauty found in the great outdoors. Communities located in native vegetation, like forests, are often referred to as Wildland-Urban Interface or WUI communities that have a greater risk of wildfires (U.S. Fire Administrator, 2022). A Report to the Council of Western State Foresters defines the term "wildland-urban interface" (WUI) as a place where humans and their development meet or intermix with wildland fuel. The three categories of the WUI community are interface, intermix, and occluded. The interface community exists where structures are next to the wildland fuels; development density is 3 or more per acre and emphasizes a population density of 250 or more

people per square mile. On the other hand, an intermix community exists where structures are scattered throughout a wildland area; development density ranges from several structures together to one per 40 acres and population density is between 28 to 250 people per square mile. The occluded community generally exists in a situation, often within a city, where structures abut an island of wildland fuels (e.g., park or open space); development density is usually similar to interface however area is less than 1,000 acres in size (Forest Service et al. 2001).

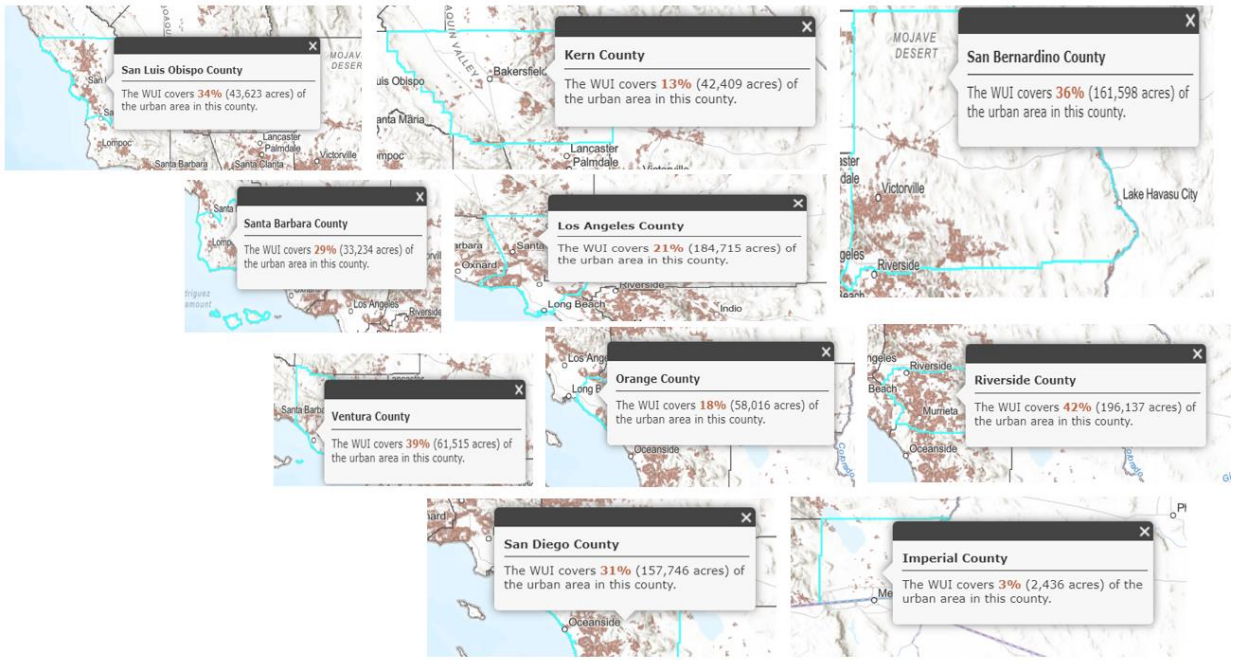
According to Gabbe, Pierce, and Oxlaj (2020), WUI is one of the most interesting areas of research on wildfires because the intersection of wildlands and urbanization represents an acute danger to humans. In November 2018, over 150,000 acres of land in Northern California were burned by Camp Fire and more than 18,000 structures were destroyed along with the loss of 86 lives. Thus, it has been categorized as the most "destructive and deadly" wildfire in California (ArcGIS n.d.). Research done by ArcGIS (n.d.) to map the WUI area per county found that the development of housing and communities within WUI of California is increasing despite the warning about the risk of wildfire based on footprints of historic fires. The Wildfire frequency and severity are more likely to increase if there is an increase in housing and the number of people in WUI areas (Gabbe et al., 2020). Rao et al. (2022) did research to identify the reason behind the rapid growth in population in these highly sensitive areas of the Southern California and they found that there is a significant contribution of timber-dependent communities and people looking for homes amid forests to the expansion of WUI areas, however, these communities are not the only responsible factors in growing WUI.

According to Li et al. (2022), in 2020, the WUI covered an area of 29,343 $km^2$ that accounted for 6.92% of the total land area in California and about 5 million housing units making it 45.13% of a total number of housings were in WUI area of CA. Figure 14 that has

been adapted from ArcGIS shows the percentage and area (in acres) of WUI in each county of southern California. The Riverside (RIV), Los Angeles (LA) and San Bernardino (SBE) and San Diego (SD) counties consist of the largest WUI areas that are 196,137 acres, 184,715 acres, 161,598 acres, and 157,746 acres respectively.

**Figure 14**

*Urban developed communities in each county classified as near or within the WUI in Southern California*




Radeloff et al. (2018) stated that the majority of wildfires in WUI areas originated from human-activity and are most difficult to control. New housing development before the wildfires complicates the job for firefighters as they must protect more houses and evacuate more residents that limit them in controlling the actual fire. Similarly, houses built after the fires are of major concern because the area burned is already at high fire risk, hence new developments in such areas are also classified into the category of fire-prone structures (Radeloff et al. 2018). NFIRS-

reported that losses in the California WUI are considerable in terms of size and severity. The average cost per year of repairing or rebuilding a damaged or destroyed building due to wildfire in WUI areas is $154.6 million in California. An additional $3 million per year is incurred in property damage due to mobile structure fires, and vehicle fire incidents in these areas add $19.4million to the losses (U.S. Fire Administrator, 2021).

According to Dennison et al. (2014), Southern California and Mediterranean California have been experiencing a reduced number of fires recently. The reason for that 2014 dip in the number of wildfires may be due to the decrease in reported fire ignitions that are originated by human-activities or due to more efforts taken into suppressing the fires (Nauslar et al., 2018). Tropical grasslands where fire is often not limited to wildland fuel or flammability, are mainly near or within WUI areas (Pausas and Ribeiro, 2013), leading to depreciation in direct correlation of wildfires in WUI areas with climate variability than in other areas of wildland with lower or no human activity is observed yet.(Nauslar et al., 2018).

Although some residents may belong to high-income groups, it cannot overshadow the fact that thousands of low-income individuals also live in fire-prone places and must be prepared for fire or recover from it without the required resources. For instance, vulnerability of many individuals living in rural areas, low-income neighborhoods and immigrant communities increases due to lack of access to the necessary resources and struggle to pay for the insurance, rebuilding or continual investment in fire safety procedures. The inequalities between two economic groups became evident after the 2017 wildfires in Sonoma County, California, where people experienced housing problems due to outrageous pricing on rentals (Davies et al. 2018). It is highly possible that low-income residents bear more risk of fires if controlling measures such as cutting trees, creating fire breaks are unreasonably expensive (Brenkert-Smith et al. 2012).

During the research on subsidized households and wildfire hazards, Gabbe et al., (2020) found that about 100 mobile home parks were destroyed by wildfires in California in 2019 alone. Income of people living in such mobile home parks is also less than that of the average household income nationwide (Gabbe et al., 2020).

Over the past decades, the fire season in Southern California has become longer and more intense, beginning in late spring (May-June) and continuing until October (Western Fire Chiefs Association, 2022). Thus, WUI residents in Southern California are exposed to the effects of wildfire for almost the entire year (Roberson, 2017). According to Congressional Research Service (2022), an average of 89% fires were caused by human activity during a time span of 4 years between 2017 and 2021. In 2017 and 2018, California experienced some of the most deadly and destructive wildfires causing the loss of 150 lives, leaving behind thousands of damaged homes, and unhealthy air affecting the health of millions of urban and rural Californians. (Kerrigan, 2020). Research from Syphard et al., (2017) presents an increase in fire activities as a result of climate change coupled with housing development within and adjacent to wildland areas i.e., WUI. San Diego county is experiencing a trend of enormous expansion of low to medium density housing in wildland areas. Thus, it has been enforcing fire codes for building constructions in WUI since 1997; that were revised in 2004 and 2008 after the fire events of 2003 and 2007 Syphard stated. The two most important factors that affect structure survival are structure density and structure age at landscape scale. That means, young structures are more likely to survive wildfires in higher-density areas compared to the lower-density areas (Syphard et al., 2017).

Research done on how housing arrangement and location determine the likelihood of housing loss due to wildfire by Syphard et al.,2012 indicates that 4% of 687,869 structures were

located within one of 40 fire perimeters in San Diego county and in the fires that occurred from 2001 to 2009, more than four thousand structures were destroyed and an additional 935 were damaged. Another study by Mueller et al., (2009), found that the loss of 24 lives and over 750,000 acres of land burned due to 14 wildfires in five counties of Southern California over the period of just 13 days starting from Oct 21$^{st}$ to Nov 3$^{rd}$ in the year 2003. In the same study Mueller et al. found that in Southern California, after the first wildfire, the price of houses declined by 10% and after the second wildfire, there is again a price drop of 23%. The government provides subsidies for living in such high-risk areas along with financial aid to the victims of natural disasters, Mueller stated. This causes both Federal and Local governments in the United States billions of dollars of losses every year for emergency funds.

**Uniqueness about Research**

While population and housing development are increasing in the WUI area of Southern California, very few households have the knowledge about the characteristics of these homes and therefore can avoid situations that can cause and spread the fire. Overall, the reasons behind the growth in population in the WUI area along with the implications of wildfires on the lifestyle or health of people living in nearby areas have been studied already. However, there has been no evidence published that can explain if the frequency and severity of wildfires have any association with the increasing WUI area and its resident population in Southern California. Similarly, there has been a lack of research on being able to predict wildfire occurrences in regard to the amount of carbon emission present in a certain area or region. Weather creates another complicated variable. The goal of predicting wildfire frequency and severity using these variables is unique and difficult. However, if attained the rewards will be great.

## Chapter 3 - Methodology

Various methods are available to be used to research how carbon emission, weather, and population impact the frequency and severity of wildfires in Southern California. Several datasets obtained from different sources are merged to form a data mart that includes wildfire data, carbon emission data, weather data, and population data. Exploratory data analysis and models such as linear and logistic regressions, XGBoost, and map overlays will be used to gather results, however, methods used to collect the necessary data are more complex. Each variable, starting with the dependent variable that is wildfires, and the subsequent independent variables: population, weather, and carbon emissions, had specific methods of data collection. These methods are explained in the individual section for each component.

### Data Sources and Collection Methods

#### *Wildfire Data*

The wildfire data was collected from the CalFire website as a comma-separated values (CSV) file that contained 2062 wildfire incidents recorded during the calendar years of 2013 through 2022 with the latest update made on October 21, 2022 (2022 Fire Season Outlook, n.d.). This wildfire dataset contained detailed information on the incidents recorded throughout the past decade, such as the wildfire incident unique identifier number, name, incident start and end dates in UTC time format, and the total acreage burned. The dataset also provided detailed information regarding the wildfire's location, including the county it took place in, the address, longitude, and latitude for precise geolocation mapping. The location variables will be utilized to merge this dataset with the collected weather, carbon emission, and population datasets. Other variables present in this dataset incident report, such as the fire department unit responding to the incident, were deemed irrelevant and were excluded from the analysis. This wildfire dataset was

then filtered to keep only the wildfires recorded in counties defined as the Southern California region in this research.

For the most comprehensive wildfire data collection, historical wildfire data was also collected from the National Interagency Fire Center (NIFC) which provided 243,727 wildfire data, which was prepared by the Wildland Fire Interagency Geospatial Services (WFIGS) (WFIGS - Wildland Fire Locations Full History, n.d.). The WFIGS, hosted on the NIFC site, reports that the dataset contains all wildland fire incidents collected by the Integrated Reporting of Wildland Fire Information (IRWIN) services, however, this ongoing data repository project may contain incomplete data on wildfires that occurred prior to the year 2014. This dataset was last updated on April 28, 2022, thus excluding the wildfires that occurred in the past seven months. As WFIGS reports, the wildland fires included in this dataset were categorized as Wildfire (WF), Prescribed Fire (RX), or Incident Complex (CX) based on the nature of the fire. In order to accurately select the wildfire incidents, only fire records tagged as WF were selected for this research analysis. Furthermore, the state and county attributes were used to select a subset of the data that included only wildfires reported in the counties that were defined as the Southern California region in this research. Similar to the dataset collected from the CalFire website (2022 Fire Season Outlook, n.d.), the WFIGS dataset contains attributes such as the fire incident unique identifier number, county of the incident, and longitude and latitude of the initial reporting site. However, unlike the CalFire dataset, the WFIGS dataset did not report the total acreage burned, instead it reported the estimated number of acres burned upon the fire discovery (Discovery Acres) and the estimated number of acres burned daily (Daily Acres) (WFIGS - Wildland Fire Locations Full History, n.d.). Thus, these parameters were used along with the total duration of the fire incident using the start and end dates of each recorded incident to

calculate the total acreage burned. Other variables including the fire code, the dispatch center ID, and fire behavior characteristics were deemed irrelevant and were eliminated from this research analysis.

***Carbon Emissions Data***

Carbon Emission data was collected from various sources with different characteristics. Several data files were downloaded in an excel format from the US Carbon Monitor. The Carbon Monitor itself is a frequently updated daily $CO_2$ emission dataset, created by researchers and professors in order to monitor the variations of $CO_2$ emissions from fossil fuel combustion and cement production since January 1$^{st}$, 2019, at the national level with near-global coverage (Lui et al. 2022). Daily carbon emissions are estimated from a diverse range of activity data, such as hourly to daily electrical power generation, monthly production data and production indices of industry processes, daily mobility data and mobility indices of road transportation, individual flight location data, and monthly data, and monthly fuel consumption data.

The Carbon Monitor data presents the dynamic nature of $CO_2$ emissions through daily, weekly and seasonal variations as influenced by workdays and holidays, as well as the unfolding impacts of the COVID-19 pandemic (Lui et al. 2022). The dataset being utilized for this study was obtained from the US Carbon Monitor, a subset of the Carbon Monitor and includes daily metric tons of carbon dioxide used for the state of California by each sector: Power, Ground Transport, Industry, Residential, Domestic Aviation as well as International Aviation, for the year 2020 to 2021.

A lot more carbon emission data was gathered from the U.S. Energy Information Administration, including State energy-related carbon dioxide emissions by year from 2005 to 2016, unadjusted and adjusted. This dataset measures the carbon emission in million metric tons

for units and was downloaded in CSV format. It is to be noted that during the 2015-2016 period, national emissions decreased by almost 2%. And due to differences in how the national and state data sets are calculated, the total for all states is not the same as the total for the United States (U.S. Energy Information Administration, 2019). Earlier in 2019, an adjustment factor was introduced to match the total for the United States, this factor was then distributed to the states in proportion to each state's share of the total: making the dataset to be referred to as either adjusted or unadjusted carbon emission values for each sector.

The U.S. Energy Information Administration (2022) also portrayed carbon emission datasets by different plants and regions for different gasses, such as $CO_2$, $SO_2$, and NOx for the years 2013 to 2020. This was downloaded as separate files for each year, all available in excel format. The dataset is also separated by plants and regions, rather than having them in the same file. This introduces another category differentiating carbon emissions, depending on the filters being observed at a given time. Each of the files contains data points with the above descriptors along with these further classifications: State, Plant Name, Sector Group, Sector Code, Fuel Code, Fuel Consumption, Emissions measuring metrics, Balancing Authority Information, and Other consumption units.

Carbon emission rates from the years 1959 up to 2021 were also collected through Statista, in an excel file format. The atmospheric level of carbon emissions has been rising ever so steadily, and in 2021, reached a high of 416.45 parts per million, in comparison to 1960 levels which stood at about 316.91 parts per million (Tiseo, 2022). Carbon dioxide emissions largely come from human activities such as burning fossil fuels and deforestation are primary drivers of climate change as well as ever-growing wildfires, however further analysis is needed to determine the impact of carbon emissions on the above-mentioned.

***Weather Data***

For the primary data collection of weather, the National Oceanic and Atmospheric Administration (NOAA) was used heavily. The NOAA is an American scientific and regulatory agency within the United States Department of Commerce that forecasts weather monitors oceanic and atmospheric conditions, charts the seas, conducts deep sea exploration, and manages fishing and protection of marine mammals and endangered species in the U.S. exclusive economic zone. Information on weather cycles such as the North Pacific Jet Stream (NPJ) were valuable data points.

Other means of finding data were sources from scholarly articles covering different weather phenomena such as the Palmer Drought Severity Index (PDSI) covered in an article from the National Center for Atmospheric Research. The PDSI was originally developed by Palmer with the intent to measure the cumulative departure in surface water balance. It incorporates antecedent and current moisture supply and demand into a hydrological accounting system that includes a 2-layer bucket-type model for soil moisture calculations.

The National Integrated Drought Information System (NIDIS) is partnered with NOAA and provides detailed map overlays of weather conditions. These weather conditions include precipitation, temperature, humidity, drought level, and wind direction/severity. These weather conditions are depicted over a map with colors indicating different levels of severity. This method is useful when explaining differences in areas over a span of land.

Other published scholarly research includes information on the Santa Ana winds of California. The Santa Ana wind is a hot, dry, foehn-type, easterly, or northeasterly wind that blows from the deserts east of the Sierra Nevada to the coast of southern California. It tends to occur in winter and spring. While it is named after the pass and river valley of Santa Ana,

California, it can affect much of the southern California region (Raphael, 2003). These winds are attributed to bringing either rain to quell wildfires or hot dry air to aid wildfires.

The University of Nebraska-Lincoln is partnered with the National Drought Mitigation Center NDMC) providing more map overlays as well as data sets to use in research. The data is available in CSV format and can quickly be uploaded and exploited. The datasets include temperature, fires, precipitation, and soil moisture to name a few. The datasets can be specified by location, file format, data type, and region.

Lastly, for weather, the use of past weather by zip code program was used. The Global Historical Climatology Network (GHCN) provides the service utilizing data sets from NOAA. This feature allows the creation of tailored data sets in excel or CSV formats. The program enables a user to check what the weather was like on specific dates in history. This program can make specific location data easy to attain to answer specific questions.

### *Population Data*

Datasets collected on populations consist of two categories: the data representing resident population and income of southern California, and the data representing population density, housing units, and area covered by WUI in Southern California.

**Resident Population and Income.** Resident population and per capita personal income data was collected from Federal Reserve Economic Data (FRED). The data was divided into multiple CSV files: one for each county in the Southern California region that includes Imperial, Kern, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, San Luis Obispo, and Ventura. Observations in these data files were recorded annually for 5 decades from 1971 to 2021. The unit of the resident population is thousands of persons, and the value of per capita personal income is stored as dollars. A new variable called Year has been added to

datasets to identify the year of each observation, and the data from several CSV files have been combined into one data file at the county level.

Census data for total population and poverty in Southern California counties have been compiled in a single data file. This data, obtained from B01003 and S1701 tables of 1-year estimates available at the American Community Survey (ACS) 2021, includes tract-level total population count and the number of people for whom status is determined as below the poverty level (U.S. Census Bureau, 2021 A). The data was available in separate CSV files for each year from 2010 to 2020 and was combined at the county level to form a single data file containing information about the total population, number, and percentage of people for whom poverty status has been determined.

Demographics and housing data for Southern California are derived from the DP05 table of ACS 2021 that provides estimates of urban and rural populations, housing units, and characteristics reflecting the boundaries of urban areas that are defined based on Census 2010 data. All the data from ACS 2021 has been collected for the period of 12 years from 2010 to 2021, and several files containing actual data and metadata were compressed before extracting from the U.S. Census Bureau(2021 B) website. The Census Bureau updates the estimates approximately every year; however, the data for the year 2022 has not been released yet.

Datasets collected from California Health and Human Services (CHHS) Open Data Portal contains data that provides the number of families with incomes below the living wage for counties and regions in California. According to CHHS (2020 B), the living wage is the basic needed budget of a family after adding all relevant taxes; it does not include publicly provided income or housing assistance. CHHS (2020 B) showed that the data includes the percentage of families below the living wage that has been calculated using data from the Living Wage

Calculator (Glasmeier, 2022) and the American Community Survey (U.S. Census Bureau, 2021 C). Low-income populations have disproportionately lower wages and poorer housing that may influence preventive measures taken by such households to avoid or reduce wildfires.

Another dataset obtained from CHHS(2020 A) contains data to represent the percentage of the total population living below 200% of the Federal Poverty Level (FPL), and the percentage of children living below 200% FPL for counties, cities, towns, public use microdata areas, and census tracts of California. The table includes the data that represents overall poverty from 2011-2015 and child poverty from 2012-2016 along with the classification of population based on race or ethnicity (CHHS, 2020 A).

**WUI – Population, Area, and Housing.** WUI data is from the University of Wisconsin SILVIS lab (2022) and has been widely used in the literature (Gabbe et al., 2020; Radeloff et al., 2018). The SILVIS lab (2010) assigned each census block as non-WUI, intermix WUI, or interface WUI at three points in time (1990, 2000, and 2010). The dataset includes the area of each WUI category in kilometers and the percent of the area covered in each county, along with household and population count for each census block at each of the three time periods.

Geospatial data collected from the U. S. Department of Agriculture (USDA) has supported wildland fire research, and inquiries into the effects of housing growth on wildfires (Radeloff et al., 2022). The authors claim that they integrated U.S. Census and USGS National Land Cover Data using the Geographic Information System (GIS) to map the Federal Register definition of WUI areas of the United States from 1990-2020. The information presented by these data are housing and population densities for 1990, 2000, 2010, and 2020; wildland vegetation percentages for 1992, 2001, 2011, and 2019; as well as WUI classes in 1990, 2000, 2010, and 2020 (Radeloff et al., 2022). The data provided by Radeloff et al., in 2022 is the third

edition; the first edition was provided by Martinuzzi et al. in 2015 that represented the 2010 WUI areas, housing, and population of the conterminous United States; Radeloff et al., 2017 created a new dataset, known as the second edition, that contained data to represent the same parameters of WUI from 1990-2010.

According to ArcGIS Pro 3.0 (n.d. a), geospatial data is available as a geodatabase that contains various types of geographic datasets held in a common file system folder or a multi-user relational database management system (RDBMS), such as Microsoft SQL server and Oracle. The website further states that the datasets can be accessed through either ArcGIS or a database management system using a structured query language (SQL). Essential relational database concepts are utilized in the storage model of the geodatabase, thereby providing a geodatabase with the advantages of an underlying database management system (ArcGIS Pro 3.0, n.d. b). The schema, rule, base, and spatial attribute data for each geographic dataset are stored in simple tables and well-defined attributes, and this approach provides a formal model for storing and working with the data that allows users to create, modify and query tables and their data elements using SQL (ArcGIS Pro 3.0, n.d. a, para. 1).

The core of a geodatabase consists of a standard relational database schema and includes a series of tables, column types, indexes, and other database objects. The database consists of two primary sets of tables: system tables and user-defined tables. One or more user-defined tables are used to store a dataset in a geodatabase, wherein the metadata required to implement geodatabase properties, data validation rules, and behaviors is stored and managed in system tables. According to ArcGIS Pro 3.0 (n.d. a), four main tables that contain the information related to the schema in geodatabase are:

- **GDB_Items:** Items such as feature classes, topologies, and domains within a geodatabase

are listed in this table.

- **GDB_ItemTypes:** A list of already defined item types such as Table are included in this system table.
- **GDB_ItemRelationships:** Provides an association between two items, for example, the relationship between feature classes and a feature dataset is stored in this table.
- **GDB_ItemRelationshipTypes:** list of relationship types is stored in this table, for example, Dataset in Feature Dataset.

According to ArcMap 10.8 (2021), ArcGIS software and database management systems share the responsibility for the management of geospatial data. Certain management aspects such as disk-based storage, the definition of attribute types, and processing an associative query and multi-user transaction, are assigned to the database management system, whereas the ArcGIS application is responsible to define the schema that represents various geospatial datasets and domain-specific logic so that the integrity and utility of data are maintained (ArcMap 10.8, 2021). The geospatial data collected from USDA has been accessed with the help of the ArcGIS application and oracle database management system to view the nature of the data and prepare it for further analysis by performing data cleaning operations.

**Data Cleaning**

In order to perform any form of analytics on the original source data, they need to be prepped first and foremost, essentially involving data cleaning. While working with large sets of data, containing several similar but not exactly the same variables, proper data cleaning procedures need to be used in order for the data analytics results to be able to provide real insights. Kowalewski (2020) states that IBM's recent study noted low data quality costs 3.1 trillion dollars every year in the U.S. alone. Moreover, Kowalewski found that there happen to be

several benefits to cleaning data, some of which are listed below:

- **Saved time and money** – Data cleaning with accurate data saves companies from potentially wasting both time and money, developing an effective strategy.
- **Increased productivity** – It also leads to consistent and highly functional databases, resulting in fewer errors, meaning faster, more effective workflows, which directly impacts productivity.
- **Better decision-making** – Cleaner data leads to better decisions, there is a direct correlation between clean, quality data and reliable business insights.
- **Maintained reputation** – Making decisions based on inaccurate data, makes a business look bad and unprofessional. But with useful insights, it has the opposite impact on the reputation and supports growth instead.

Kowalewski (2020), states that a typical data cleaning process involves the following 5 steps to get the data ready for analytics: Data Audit, Workflow Specification, Workflow Execution, Validation, and Reporting. The author states that, often, Data Auditing can involve "Data Profiling", which is a technique helping in examining the data to create general, informative reports about what is in the data set (Kowalewski 2020). Once the data has been cleaned and is ready for analytics, this study will be utilizing creating a data mart in order to analyze the data and see any potential patterns and trends between carbon emission, weather, population, and wildfires taking place in Southern California.

**Data Mart**

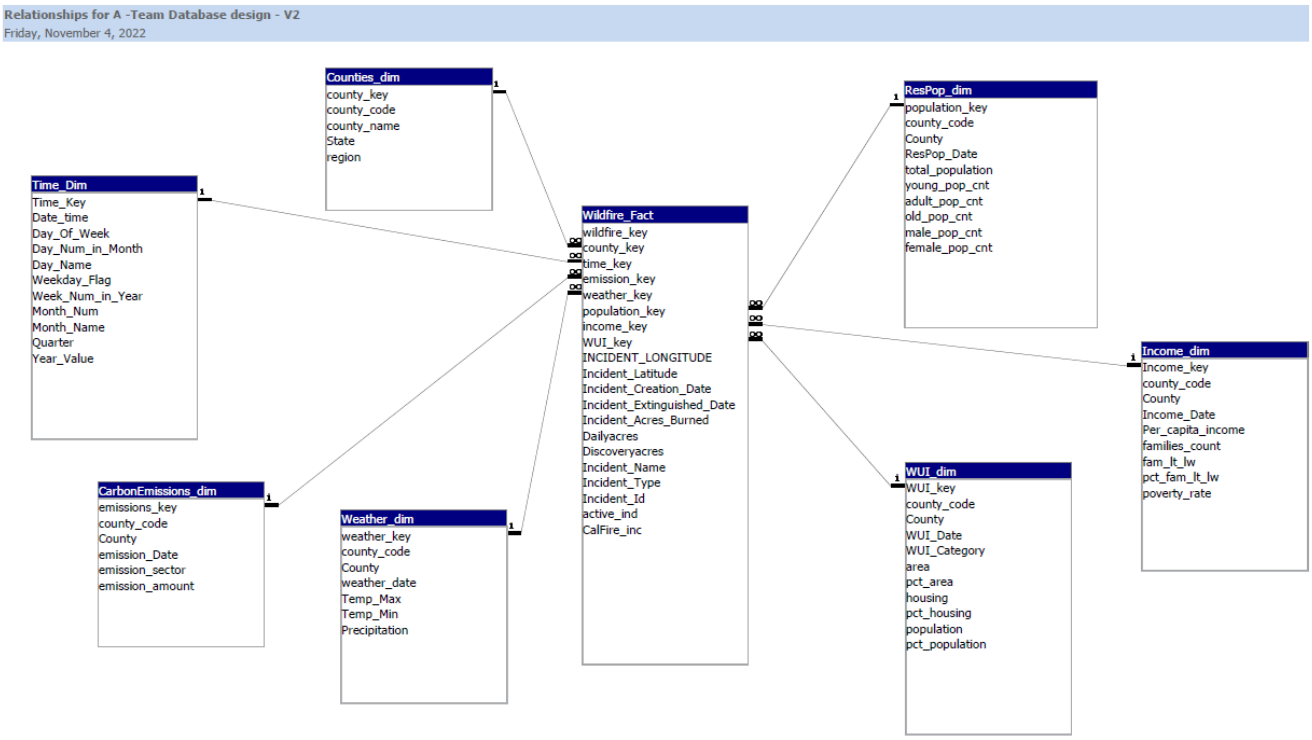
According to Coronel and Morris (2016), a data mart is defined as a data management system that supports the analysis of data to make more informed decisions. Coronel and Morris (2016) further state that the data mart contains data from multiple sources, such as relational

databases and data files, and is stored in the same way for all components, thereby providing users with a unified view. The data mart allows users to analyze historical data and provides information about one subject to help make data-driven decisions. The flow of data in a data mart is represented through time, meaning that the changes in data over time are integrated and preserved in the data mart; and to summarize and aggregate the data, a time key is assigned to each data point when the data is entered into the data mart. Because of all these features, building a data mart is preferred over an operational database when it comes to analyzing the data for decision-making purposes. Often, data marts consist of redundant data with the time that helps users with multiple views of the same information, whereas operational data is not likely to have any redundant data stored in it. The data collected for this research is multidimensional data because it contains different features and attributes i.e., multiple rows and columns; and, to map multidimensional decision-support data into relational data marts, a star-schema design technique is used (Coronel & Morris, 2016).

Four components of the dimensional model are facts, dimensions, attributes, and attribute hierarchies. Facts are quantitative measurements that represent specific activity and are stored in the fact table, wherein dimension tables contain dimensions that provide additional perspective to a given fact. The fact table is the center of the data mart and is surrounded by dimension tables to give an appearance of a star shape. Facts can be searched, filtered, or classified with the help of attributes stored in the dimension table (Coronel & Morris, 2016). Figure 15 shows the conceptual schema of the data mart designed to store the data collected from various sources for the components: wildfires, carbon emissions, weather, and populations.

**Figure 15**

*Data mart design*




The conceptual schema of the data mart consists of one generalized fact table and seven dimension tables, containing the data captured after an incident is considered closed, and facts have been recorded. As shown in Figure 15, the Wildfire_Fact table contains the attributes wildfire_frequency and acres_burned that are numeric variables to store measurements of wildfires in Southern California. Attributes active_ind and CalFire_Inc of the same table are binary variables and indicate whether the wildfires are active and reported to CalFire. The wildfire_Key is a primary key that is a unique identifier for each incident on Cal Fire's historical record in the table, and the other attributes are county_key, time_key, emissions_key, weather_key, population_key, income_key, and WUI_key that provide additional perspective to the fact being presented by each record in the Wildfire_Fact table. The attributes with 'key' in

their name are foreign keys from respective dimension tables to access additional information associated with the incidents log. Each dimension table has the keyword ‘dim’ at the end of the table name, and the Wildfire_Fact table represents a many-to-one (M:1) relationship with each dimension table. Each table contains a default value record as the first entry to be used whenever the join is considered an option or when information is unknown.

Dimension tables identified from the collected data are CarbonEmissions_dim, Weather_dim, WUI_dim, Income_dim, ResPop_dim, Counties_dim, and Time_dim. All the foreign keys in the fact table are primary keys in the respective dimension table; county_key is the primary key of the Counties_dim table with number as data type, that a unique number is generated for each record inserted in the table. Similarly, remaining pairs of dimension table and primary key are: Time_dim - time_key which is mandatory to indicate the exact date when the incident was defined and known by an official name based upon the point of origin; CarbonEmissions_dim - emissions_key; Weather_dim - weather_key; ResPop_dim - population_key; WUI_dim – WUI_key; Income_dim – Income_key. Counties_dim has been created to store details about counties, such as county_name, state, and region because the data collected for different components will be combined at a county level before being stored in the data mart. Time_dim is required to store details about the date and time when dimensional modeling is used to design the conceptual schema of the data mart. Each record in this diverse conceptual fact table is assigned a Time_key that provides the timestamp of the incident record. Attributes of CarbonEmissions_dim table such as county, emission_date, emission_sector, and emission_amount will store a series of new records to log any significant new data about the amount of carbon emitted by different sectors within each county of Southern California at a specific given time. For simplicity, only the most relevant dim is referenced. These may be

relevant to accuracy of information about any incidents which span geographical areas. Attributes in Weather_dim will store external references to measurements of each factor of the weather such as temperature, humidity, and precipitation, which may change during a long running wildfire incident. Data representing areas, housing units, and population count in each category of WUI will be stored in the WUI_dim table, whereas demographic information will be stored in the ResPop_dim table. Additionally, the latest known statistical record about per capita personal income, number of families with income less than a living wage, and poverty rate will be available in the Income_dim table for further analysis. All the dimension tables will contain information such as the name of counties and the date or year of observations, and may lead to a perception of redundant, for instance, denormalized data in the data mart; however, this redundancy of data will be beneficial in the improved performance of queries being executed to access the data for business intelligence purposes. The data stored in the data mart will be analyzed to answer the research question with the help of different data analysis models.

## Analysis Methods

In this study, multiple linear regression, and extreme gradient boosting (XGBoost) models will be used to test and evaluate the prediction accuracy of wildfire frequency and severity measured by the total acreage burned. The selected variables from the daily weather data, carbon emission data, and population density data will be used as input variables, known as predictor variables, when designing these prediction models. These input variables will be selected using variable selection strategies.

### *Input Variable Selection for the Models*

Since there are many variables to choose from in each category; weather, carbon emission, and population data, as predictor variables, the variable selection process becomes

crucial in the optimization of data performance and producing maximum parsimony (Shmueli et al., 2019). For the variable selection process, exploratory analysis, the forward selection, and backward elimination method will be utilized to select the combination of the input variables that would produce the best prediction model. During the explanatory analysis, the summary statistics for missing value counts, correlation tables, and boxplots will be used to ensure the selected predictor variables have a low missing value count, a strong correlation with the outcome variable, and a low number of outliers (Shmueli et al., 2019). Once the set of variables is identified as potential predictor variables, the forward selection, and backward elimination methods will be applied to select the combination of those predictor variables that would produce a prediction model with the highest $R^2$ value. With the forward selection model, predictor variables will be added one at a time and evaluated based on their contribution to the $R^2$ value of the model (Shmueli et al., 2019). On the other hand, the backward elimination method will be used to eliminate the predictors that have the least statistically significant contribution to the overall model (Shmueli et al., 2019).

***Multiple Linear Regression Model***

A multiple linear regression model will be used to depict the relationship between the numerical outcome variable such as wildfire total acreage burned and the predictor variables or input variables such as daily temperature, daily carbon emission, and population density. In simple form, the multiple regression function formula is depicted as the following:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where, Y is the outcome variable, $x_1$, $x_2$… are the predictors, $\beta_0$, $\beta_1$ … are coefficients and € is the unexplained error (Shmueli et al., 2019).

### *Extreme Gradient Boosting (XGBoost)*

Extreme gradient boosting (XGBoost) algorithm is a widely recognized machine learning algorithm that outperforms the other methods in its class (Ibrahem et al., 2021). The XGBoost algorithm implements a supervised learning approach for gradient tree boosting (Chen & Guestrin, 2016). In the gradient-boosting technique, decision trees are created sequentially (Shmueli et al., 2019). This ensemble technique of the sequential learning approach enables the model to improve its performance using the determined degree of error of the previous model (Chen & Guestrin, 2016). The uniqueness of the boosting algorithm is that each tree in this technique concentrates on the misclassified records or on the degree of error from the previous tree, consequently reducing the misclassification or error and improving the model's overall performance (Shmueli et al., 2019). XGBoost can be applied for both regression and classification problems (Chen & Guestrin, 2016), which makes it applicable to this research question considering that the total acreage burned outcome variable is a continuous variable that will be binned into three categories of low severity, medium severity, and high severity.

### *Spatial Data Analysis*

According to Chi and Zhu (2008), often, spatial effects are excluded from the analysis of population data in most sociological and demographic research. However, incorporating the spatial effects has become essential because, from a methodological standpoint, if such effects are not considered in a model but exist in the data, these effects may cause it to produce an unreliable estimation and statistical inference; for example, understating or overstating the effects of explanatory variables on the outcome variable (Chi & Zhu, 2008). In the research to understand spatial regression models for demographics analysis, Chi & Zhu (2008) state that exploratory data analysis is an important step before performing regression analysis on structured

data. Similarly, exploratory spatial data analysis (ESDA) is crucial in understanding the spatial features of the data. Visualizations generated in ESDA often display the spatial patterns of the data; spatial clusters and spatial outliers are identified by ESDA along with determining the possible ways that the statistical model might fail to represent the desired results because of spatial aspects of the data (Chi & Zhu, 2008).

**Summary of Methodology**

The already existing data sets from CalFire, U.S. Carbon Monitor, EIA, NOAA, U.S. Census Bureau, FRED, and U. S. Department of Agriculture aided this research tremendously by establishing a baseline using historical documentation. Once the baseline is established, deviations from the norm are easier to identify. The historical data will be used in models to predict the frequency and severity of wildfires in Southern California with the help of explanatory variables carbon emissions, factors in weather, population density in WUI regions, and income. Once all the data is cleaned, restructured and modified, analysis will be conducted in order to achieve the research objective. Multiple machine learning algorithms such as spatial data analysis, linear regression and predictive modeling with XGBoost will be utilized to complete this analysis.

# Chapter 4 - Results and Analysis

## Exploratory Data Analysis on individual datasets

Exploratory data analysis (EDA) was performed as part of the initial investigations to determine the pattern and anomalies in the data. Research by Telang (2021) shows that initial EDA was performed on individual datasets by itself to be able to analyze the data before making any assumptions. Telang states that an EDA can be utilized for the purpose of uncovering the underlying structure of the data, and help to determine the trends, patterns, and relationship among the data itself. Descriptive statistics and EDA were performed on the wildfire dataset of 30,023 unique records to obtain a deeper understanding of the data types, distribution, and relationships between the outcome and explanatory variables. EDA also provided detailed insight to the wildfires trend in the frequency and acreage damage observed throughout the years 2013 and 2022.

### *EDA on Variables from Wildfire Dataset*

As shown in Figure 16, the exploratory data analysis was run on the entire data set of 30,023 unique wildfire records, which revealed the distinct trend of the annual wildfire incidents recorded in the Southern California region. The *catplot* function of the *seaborn* python package was utilized to create the categorical plot, where each bar represents the total count of the unique incident_id for the given year. As shown in Figure 16, there was a distinct increase in wildfire frequency, ranging from 52 wildfires in 2013 to 7,201 in 2021. To visualize the total amount of acreage damaged by these wildfires, the Tableau data visualization tool was utilized to show the total acreage burned each year. The result is shown in Figure 17 and depicts that despite the year 2021 having the highest number of wildfires recorded during this time period, the total acreage burned is only 109,144, falling well behind its predecessor year of 698,836 acres of damage due

to the wildfires. The rolling average shows a distinct spike in the total amount of acreage burned between the years 2016 and 2017.

**Figure 16**

*The Number of Wildfires Recorded in Southern California*

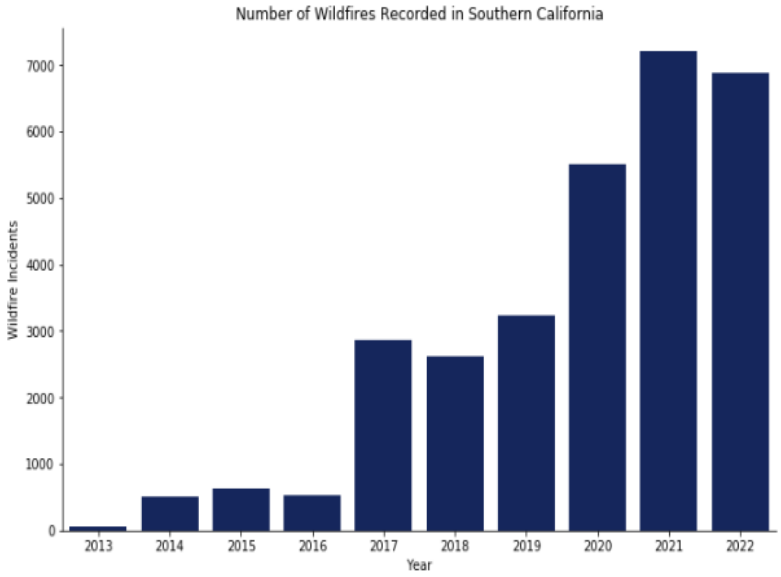
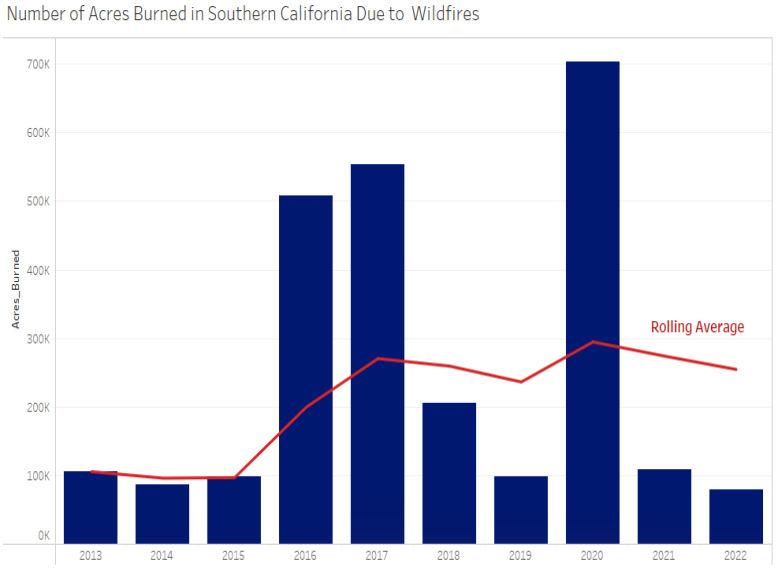



**Figure 17**

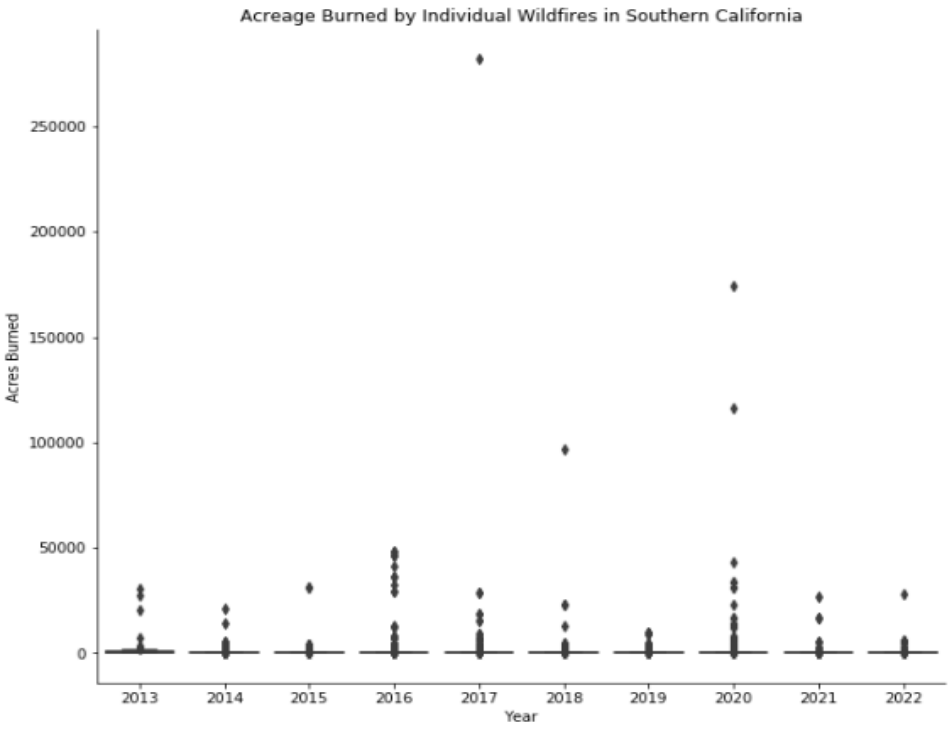
*Number of Acres Burned in Southern California Due to Wildfires*



A box plot was utilized to plot Acreage Burned by Individual Wildfires in Southern California to visualize the statistical distribution of the data points and quickly identify the potential outliers. As shown in Figure 18, the interquartile range of Acres Burned (Incident_Acres_Burned) falls between 0.01 and 0.1 acres. However, there are several distinct outliers such as the data point in the year 2017. This wildfire record represents the Thomas Fire which claimed 281,893 acres of land. Similarly, the Castle wildfire in 2020, claimed 174,178 acres. Although these magnitudes of wildfires are rare, these records were determined to be critical to this research focus and were not removed from the dataset.

**Figure 18**

*Acreage Burned by Individual Wildfires in Southern California*




Wildfire frequency was also studied across a twelve-month period to identify seasonal patterns of wildfire occurrences. As shown in Figure 19, the number of wildfires recorded per

month in Southern California had a distinct seasonal trend. The highest wildfire occurrences were observed in the months of June, July, and August, with the highest peak in July. In contrast, the highest total acreage burned (Figure 20) was observed in July, August, and December, with the highest peak in August.

**Figure 19**

*Number of Wildfires Recorded per Month in Southern California Between Years 2013 to 2022*

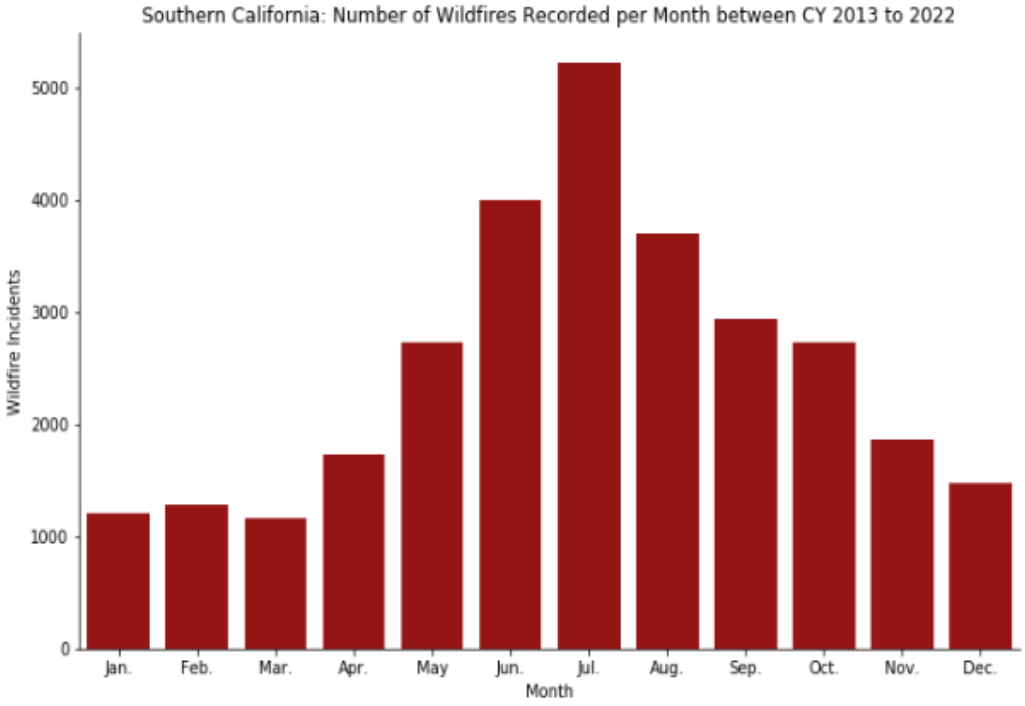
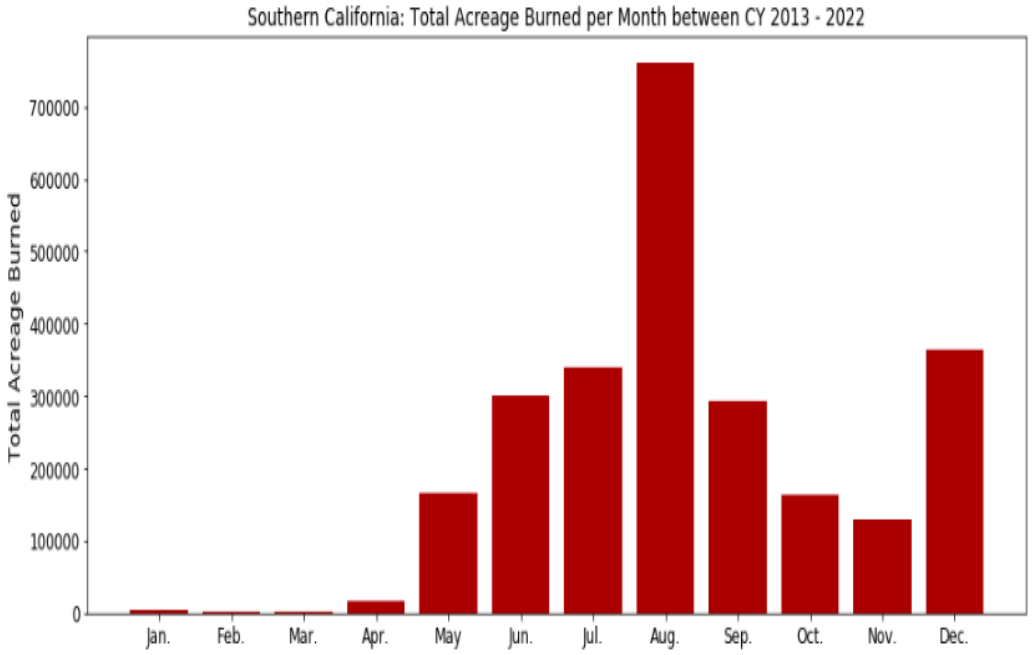


**Figure 20**

*The Total Acreage Burned per Month in Southern California Between Years 2013-2022*




The wildfire frequency and burned acreage were also studied among the Southern California counties. As shown in Figure 21, Los Angeles County experienced the highest number of wildfires within the past decade with a total of 16,207 wildfires, whereas Kern County only had 925 wildfires. However, when examining the total acreage impacted due to these wildfires, Kern County faced the highest accumulated number of acreages burned within the past decade reaching almost 465,200 acres and followed by Ventura County with 431,442 acres.

**Figure 21**

*Total Number of Wildfires Recorded in Southern California Counties Between Years 2013 and 2022*

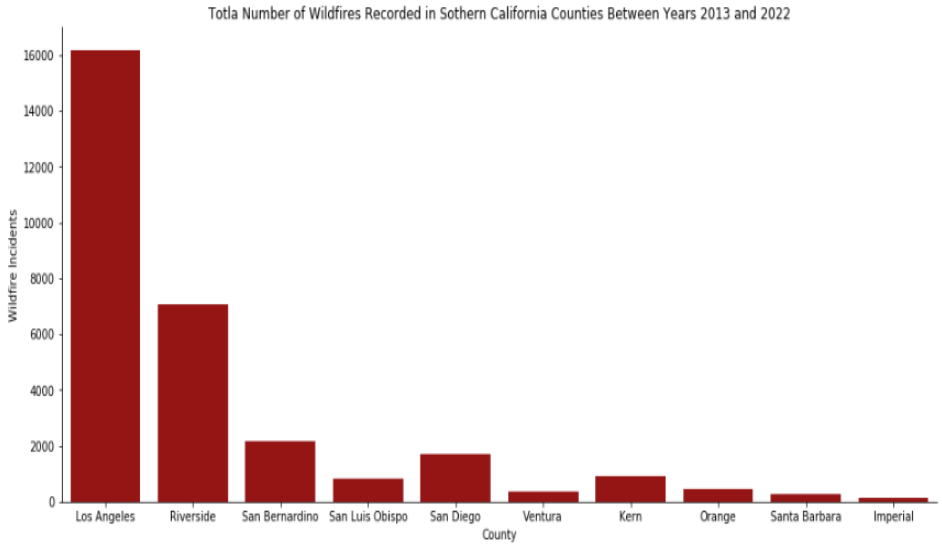
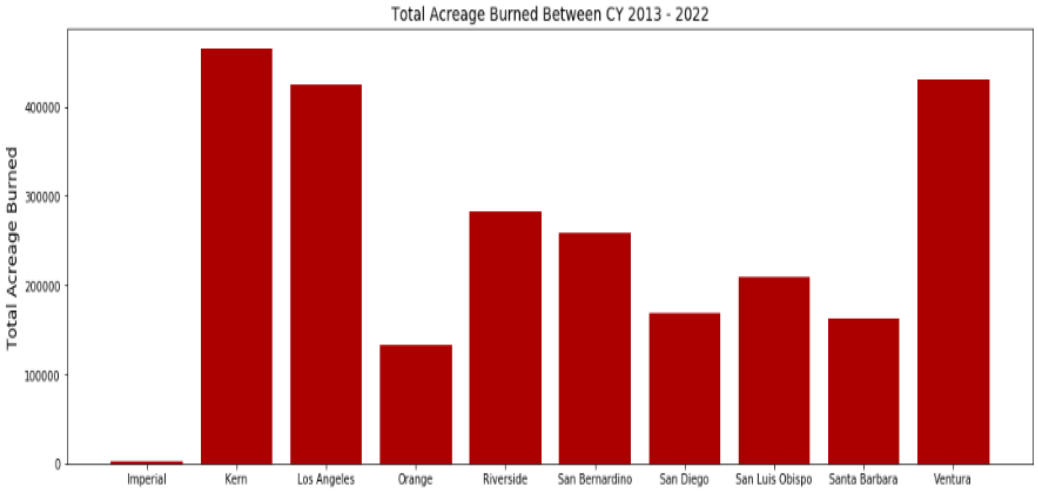



**Figure 22**

*Total Acreage Burned in Southern California Counties Between the Years 2013 and 2022*




One-way analysis variance (ANOVA) was used to explore the relationship between the outcome variable of acreage burned and individual independent variables of year, month, and

county. Each ANOVA result indicates statistical significance between the categories of the variables. The ANOVA results for acreage burned in response to year, month, and county respectively yield *p-value* < 0.05, <0.05, and 0.01. Once it was established that there is a statistical significance in the effect of independent variables and the outcome variable, a multivariate analysis of variance (MANOVA) was performed to study the effect of all the independent variables on the outcome variable. As shown in the python output of MANOVA in Table 2, all the independent variables have a statistically significant effect on the amount of burned acreage, with Wilks' lambda test statistics showing F-value = 9.17 and p = 0.000.

**Table 2**

*Multivariate Analysis of Variance for Acreage Burned*

| Intercept | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks'lambda | 0.000 | 30 | 29992 | 5.63E+17 | 0.000 |
| Pillai'strace | 1.000 | 30 | 29992 | 5.63E+17 | 0.000 |
| Hotelling-Lawleytrace | 5.63E+14 | 30 | 29992 | 5.63E+17 | 0.000 |
| Roy's greatestroot | 5.63E+14 | 30 | 29992 | 5.63E+17 | 0.000 |

| INCIDENT_ACRES_BURNED | Value | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|
| Wilks' lambda | 0.9912 | 29 | 29993 | 9.1719 | 0.000 |
| Pillai's trace | 0.0088 | 29 | 29993 | 9.1719 | 0.000 |
| Hotelling-Lawley trace | 0.0089 | 29 | 29993 | 9.1719 | 0.000 |
| Roy's greatest root | 0.0089 | 29 | 29993 | 9.1719 | 0.000 |

***Carbon Emissions And Fuel Consumption Variables***

SAS EG was applied to run Summary Statistics on 2,319 observations of the Carbon Emission dataset and was found to have no missing values; the output of this is shown below in Table 3. Having found no missing values for both the variables, Metric_Tonnes_of_CO2_Emissions and Total_Fuel_Consumption_(MMBtu), appears to be a

very advantageous aspect for further analysis being conducted on the dataset. It suggests that the dataset avoided introducing bias into the analysis. According to the Korean Journal of Anesthesiology (2022), missing data can present invalid results and reduce the competency of the analysis (as cited in How to Deal with Missing Data? 2022). The mean of carbon emission amount (Metric_Tonnes_of_CO2_Emissions) is 104,901.47, in comparison to the mean of fuel consumption (Total_Fuel_Consumption_(MMBtu)), which is 1,988,453.62, and their Kurtosis measures 21.54 and 22.96, respectively.

**Table 3**

*Summary statistics of variables Metric_Tonnes_of_CO2_Emissions And Total_Fuel_Consumption_(MMBtu)*

| Variable | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| Metric_Tonnes_of_CO2_Emissions | 104901.47 | 280009.53 | 0 | 2502355.57 | 2319 | 0 | 78405337954 | 4.36 | 21.54 |
| Total_Fuel_Consumption_(MMBtu) | 1988453.62 | 5144590.14 | 0 | 47151552.00 | 2319 | 0 | 26466808000000 | 4.45 | 22.96 |

**Table 4**

*Summary statistics of variables Metric_Tonnes_of_CO2_Emissions And Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable*

| Sector_Group | N Obs | Variable | Mean | Std Dev | Min | Max | N | N Miss | Std Error | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| COMMERCIAL | 367 | Total_Fuel_Consumption | 459662.28 | 1484074.23 | 0 | 26397982.00 | 367 | 0 | 77468.04 | 2202476300000 | 14.77 | 256.29 |
| | | Metric_Tonnes_of_CO2_Emissions | 24389.87 | 78756.88 | 0 | 1400954.00 | 367 | 0 | 4111.08 | 6202645688 | 14.78 | 256.34 |
| ELECTRIC POWER | 1553 | Total_Fuel_Consumption | 2243351.39 | 5693250.82 | 0 | 47151552.00 | 1553 | 0 | 144468.97 | 32413105000000 | 4.08 | 19.14 |
| | | Metric_Tonnes_of_CO2_Emissions | 115162.84 | 303285.13 | 0 | 2502355.57 | 1553 | 0 | 7696.01 | 91981869582 | 4.07 | 19.01 |
| INDUSTRIAL | 399 | Total_Fuel_Consumption | 2402514.26 | 4810935.95 | 0 | 30242183.00 | 399 | 0 | 240848.05 | 23145105000000 | 3.96 | 17.04 |
| | | Metric_Tonnes_of_CO2_Emissions | 139016.35 | 291033.14 | 0 | 1604967.21 | 399 | 0 | 14569.88 | 84700290695 | 3.66 | 13.19 |

After the initial Summary Statistics function was run using SAS EG, Sector_Group was used as a classifying variable and the analysis was run again to better understand the distribution of the data. Table 4 presents us with the Summary Statistics for 367 observations for the Commercial Sector, 1,553 for the Electric Power Sector, and 399 for the Industrial Sector, concerning the carbon emission and fuel consumption variables.

A PROC FREQ statement was run to understand the distribution of the data within each of the 10 Southern California counties (see Table 5) using the County variable. The 10 counties exhibit the frequency percentage in the table provided, with Los Angeles, San Diego and Kern comprising the majority.

**Tables 5**

*PROC FREQ of all variables within the Carbon Emission dataset*

**(a)**

| County | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| Imperial | 224 | 9.66 | 224 | 9.66 |
| Kern | 364 | 15.70 | 588 | 25.36 |
| Los Angeles | 629 | 27.12 | 1217 | 52.48 |
| Orange | 125 | 5.39 | 1342 | 57.87 |
| Riverside | 133 | 5.74 | 1475 | 63.61 |
| San Bernardino | 269 | 11.60 | 1744 | 75.20 |
| San Diego | 461 | 19.88 | 2205 | 95.08 |
| San Luis Obispo | 12 | 0.52 | 2217 | 95.60 |
| Santa Barbara | 17 | 0.73 | 2234 | 96.33 |
| Ventura | 85 | 3.67 | 2319 | 100.00 |

**(b)**

| Sector_Group | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| COMMERCIAL | 367 | 15.83 | 367 | 15.83 |
| ELECTRIC POWER | 1553 | 66.97 | 1920 | 82.79 |
| INDUSTRIAL | 399 | 17.21 | 2319 | 100.00 |

SAS EG was utilized in creating histograms to display frequency distribution for the carbon emission variables. Figure 23, and Figure 24, as shown below, both portray a positively

skewed unimodal normal distribution for the Metric_Tonnes_of_CO2_Emissions and Total_Fuel_Consumption_(MMBtu) variables.

**Figure 23**

*Distribution of Metric_Tonnes_of_CO2_Emissions.*

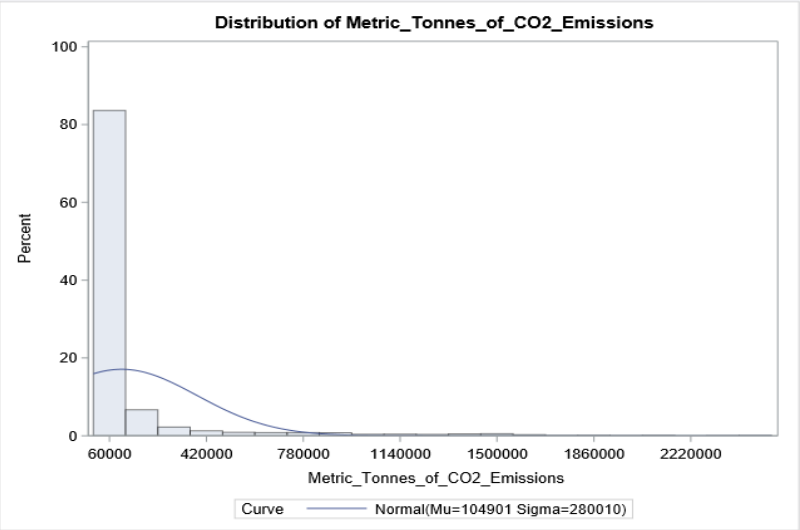
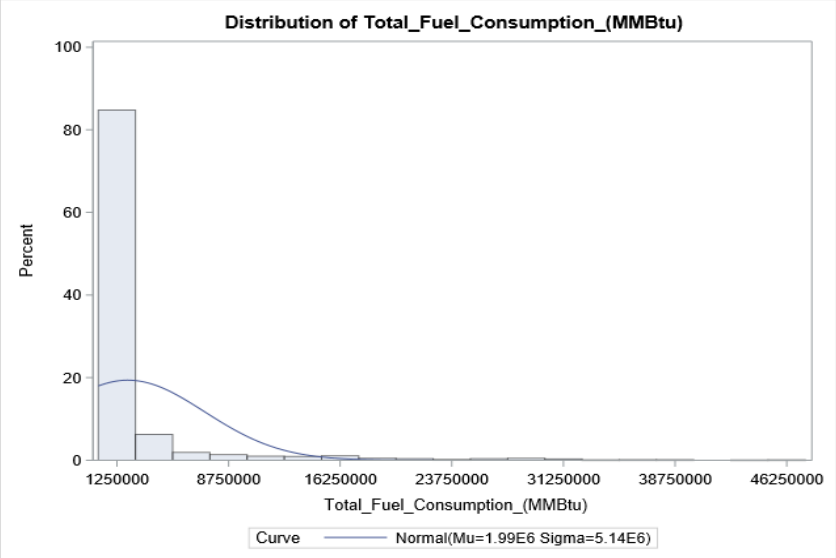



**Figure 24**

*Distribution of Total_Fuel_Consumption_(MMBtu).*



Histograms were once again created in order to better understand the frequency distribution of the above-mentioned carbon emission variables; however, this time they were created using the Sector_Group as a classification variable. Figure 25 depicts the distribution of the carbon emission amount variable with the Commercial and Electric Power sector group, whereas Figure 26 depicts the distribution for the same variable with the Industrial sector group. All of these histograms are positively skewed and show a unimodal normal distribution, however, there is a clear portrayal of higher frequency regarding the carbon emission amount variable within the Industrial sector group. Similarly Figure 27 and Figure 28 present the frequency distribution while using the Sector_Group as a classifying variable, with the exception of it being for the fuel consumption variable, once again seen portraying a higher frequency for the Industrial sector group.

**Figure 25**

*Distribution of Metric_Tonnes_of_CO2_Emissions using Sector_Group as classification variable.*

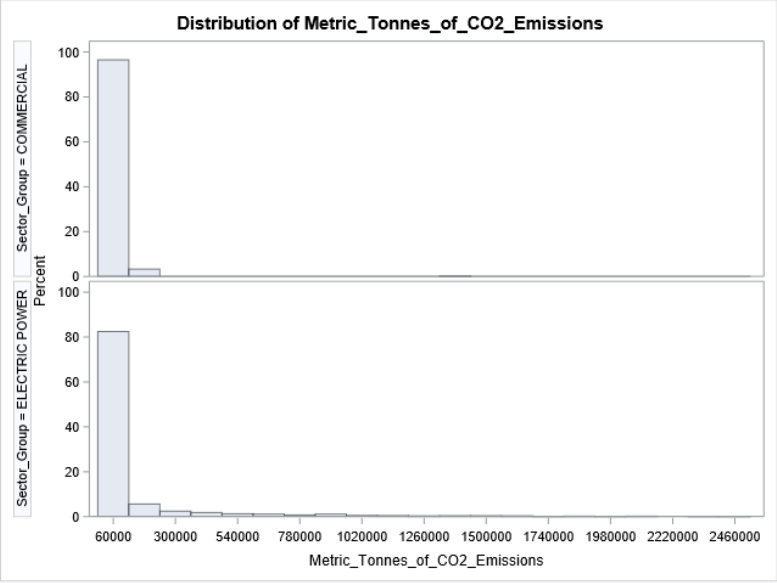


**Figure 26**

*Distribution of Metric_Tonnes_of_CO2_Emissions using Sector_Group as classification variable.*

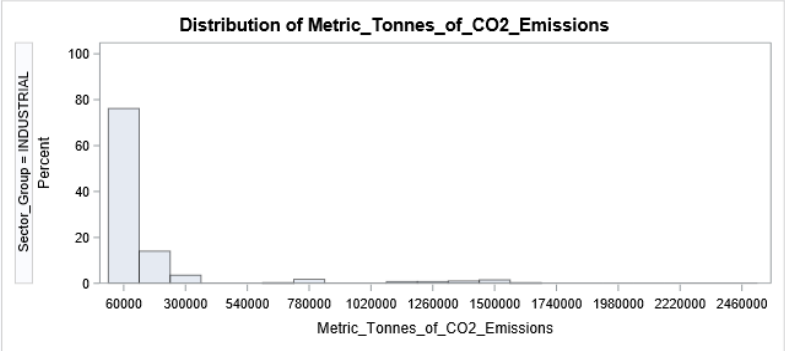



**Figure 27**

*Distribution of Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable.*

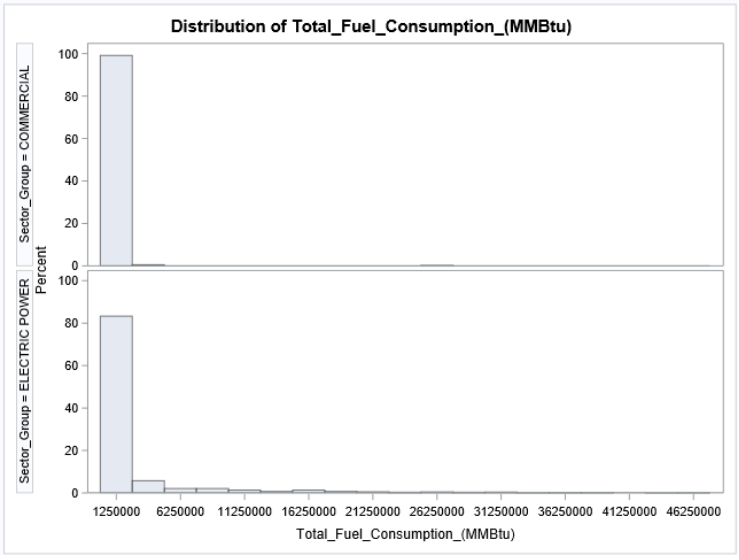
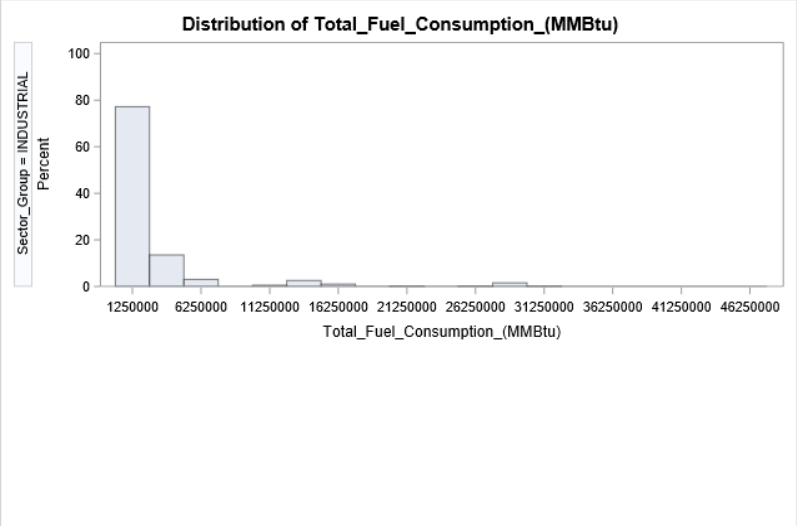


**Figure 28**

*Distribution of Total_Fuel_Consumption_(MMBtu) using Sector_Group as classification variable.*




The weather data consisted of three major variables: PRECIPITATION containing values of average daily precipitation, TEMP_MAX with an average of maximum temperature, and TEMP_MIN containing the average value of minimum temperature for each county per month.

***Precipitation***

The average precipitation count displayed in Figure 29 shows that the most common average precipitation amount is close to 0 mm a month for Southern California. This aligns with the assumption that lack of precipitation is leading to permissive environments for wildfires. Figure 29 is positively skewed with high kurtosis, and the descriptive statistics shows that the mean average temperature is 0.88 mm per month from 2013 to 2020 indicating the majority of data points around 0. While there are data points all the way up to 6mm a year they are not the norm and form the tail of the skewed histogram. All these points together indicate that lack of rain may impact the frequency and the severity of wildfires in Southern California.

**Figure 29**

*Precipitation histogram with descriptive statistics*

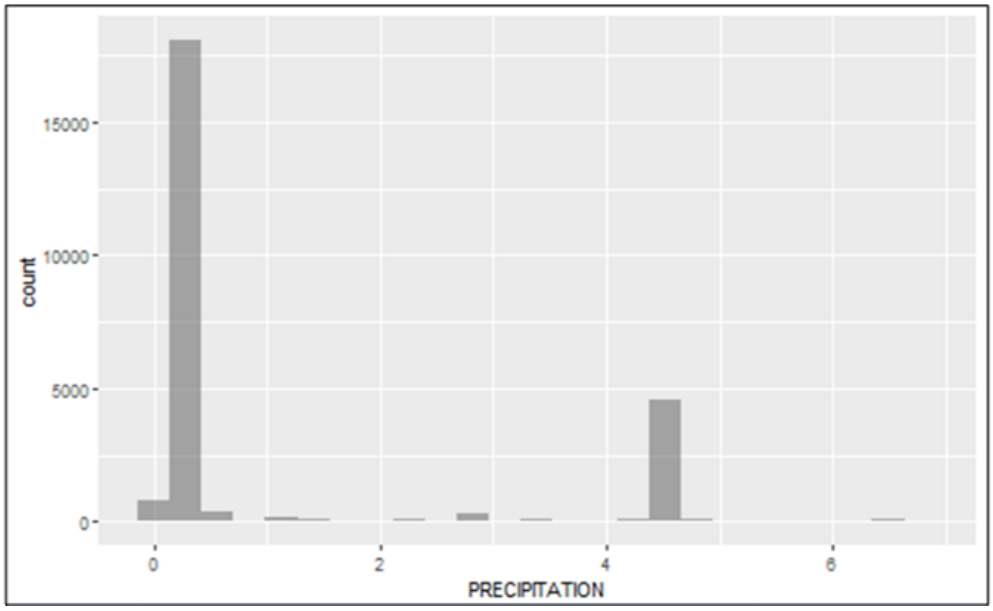



**Table 6**

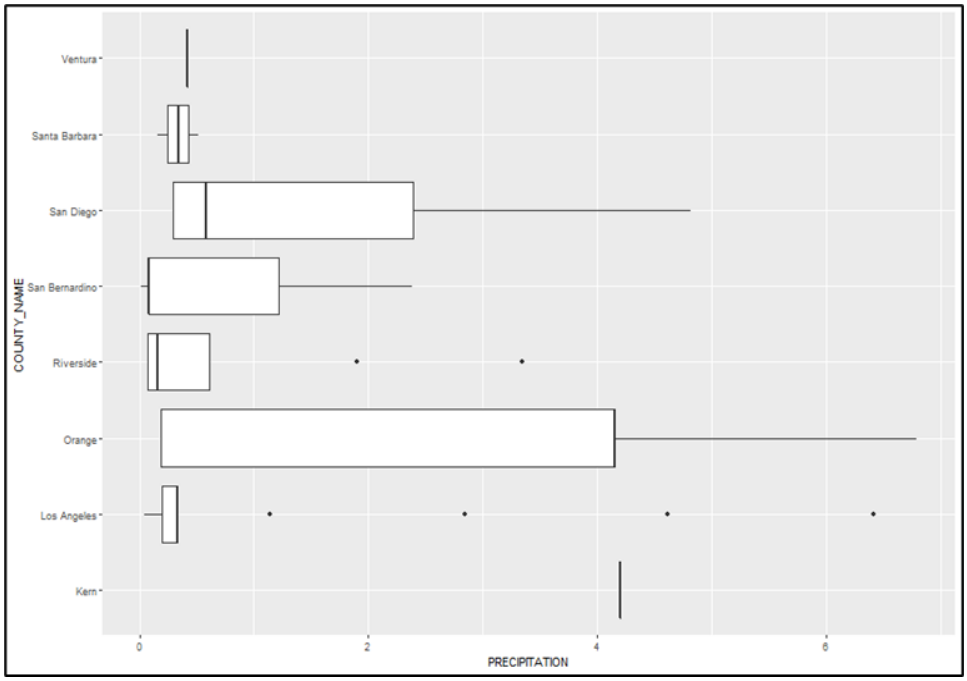
*Descriptive Statistics of Precipitation*

| Minimum | Q1 | Median | Q3 | Maximum | Mean | SD | N | Missing |
|---|---|---|---|---|---|---|---|---|
| 0.012 | 0.19 | 0.33 | 0.33 | 6.78 | 1.18 | 1.17 | 24905.00 | 0 |

The boxplot in Figure 30 shows the amount of average precipitation with respect to each county in Southern California and depicts that the amount of precipitation in Imperial County was zero from 2013 to 2020. Although Orange County shows the highest amount of precipitation for the same period, it was below 4 mm per month. The boxplot also indicates that counties such as Ventura have become dryer than others because there is a lack of precipitation falling on the soil and foliage to make damp conditions thereby increasing the severity and frequency of wildfires in the Southern California area.

**Figure 30**

*Box plot showing precipitation amount by County*




***Maximum Temperature***

Considering the level of precipitation in Southern California the maximum temperature made the possibility of a dryer environment even worse. The histogram in Figure 31 shows that the mean of maximum temperature is 17 degrees Celsius from 2013 to 2020. However, the greatest number of occurrences for maximum temperature is closer to 18 degrees Celsius. The mean of maximum temperature increased after the final analysis on the data excluding outliers, indicating a higher temperature with nearly 0 mm of rainfall per month in the Southern California region from 2013 to 2020.

**Figure 31**

*Histogram of Maximum Temperature*

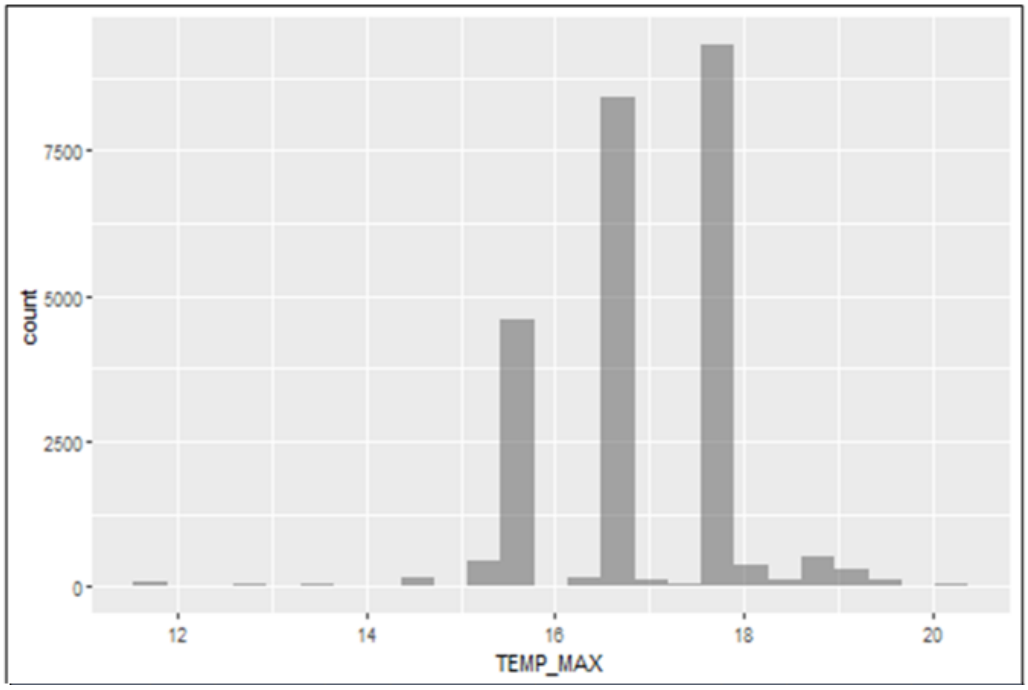



**Table 7**

*Descriptive Statistics of Maximum Temperature*

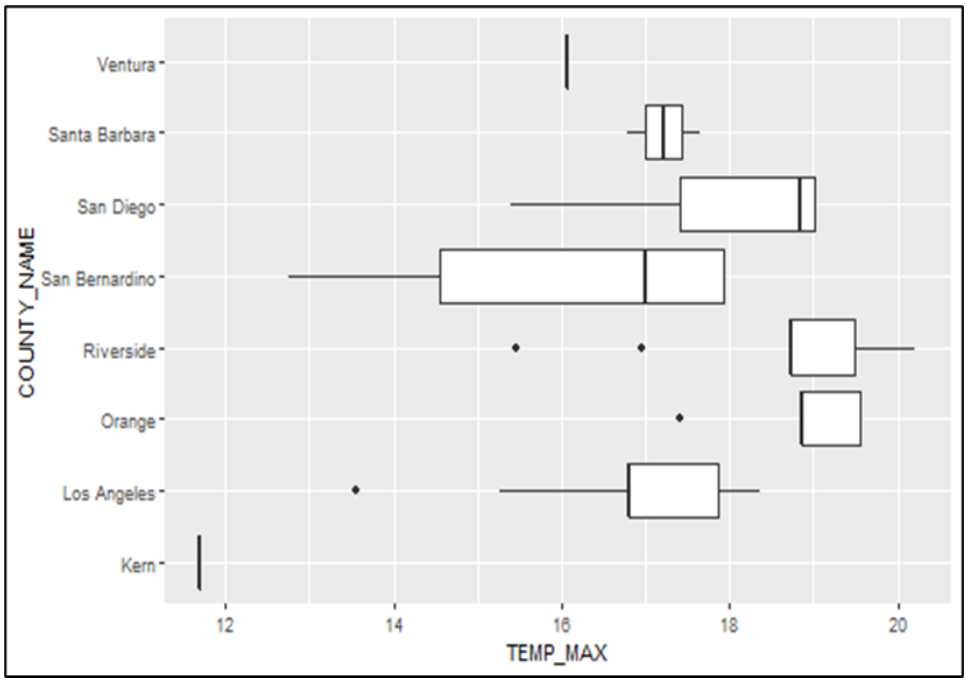
| Minimum | Q1 | Median | Q3 | Maximum | Mean | SD | N | Missing |
|---|---|---|---|---|---|---|---|---|
| 11.7 | 16.8 | 16.8 | 17.8 | 20.21 | 16.9 | 1.06 | 24905.00 | 0 |

When looking at the maximum temperature associated with the Southern California region, Figure 32 shows a comb style distribution. Riverside and Orange County had the highest maximum average temperature. While Kern County had the lowest average maximum average temperature. This data showed that the Kern County which had the highest value of precipitation also had the lowest value of maximum temperature for the same time duration; on the other hand, Ventura County had the lowest precipitation and 16 degrees Celsius as an average of maximum temperature. Riverside had the highest maximum temperature average and one of the

lowest precipitation averages. Meaning from the expository data analysis, Riverside County was the driest and hottest region in Southern California.

**Figure 32**

*Bar chart showing temperature maximum by County*




***Minimum Temperature***

After reviewing both precipitation average and maximum temperature average, the next variable to explore was minimum temperature average. This was an important variable because, if there was precipitation present and the temperature was too cold, then the precipitation would freeze on the surface of the ground rather than permeate into the soil and quickly evaporate into the atmosphere. Figure 33 showed  a bimodal distribution focused around both 5 and 7 degrees Celsius. According to the descriptive statistics the mean average temperature minimum is 5.8 degrees Celsius. This temperature was cold but not freezing meaning some of the precipitation would be able to permeate into the ground as liquid.

**Figure 33**

*Histogram of Minimum Temperature*

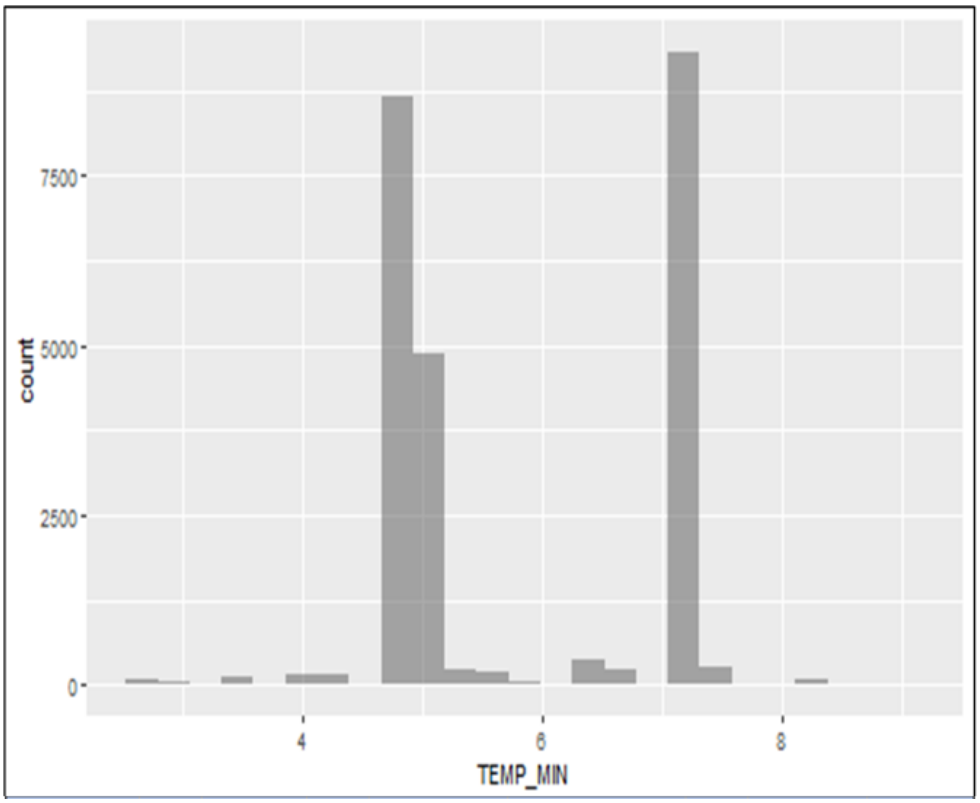



**Table 8**

*Descriptive Statistics of Minimum Temperature*

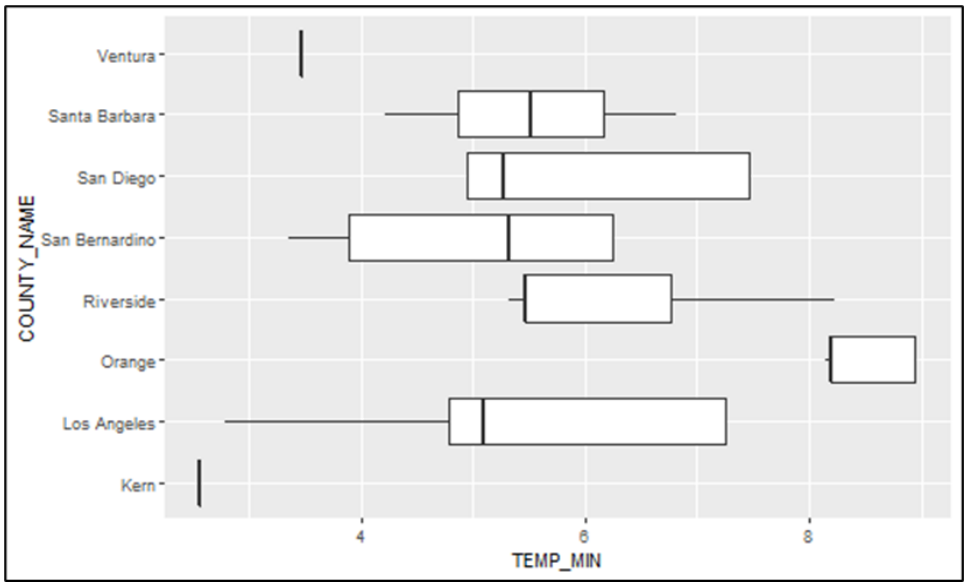
| Minimum | Q1 | Median | Q3 | Maximum | Mean | SD | N | Missing |
|---|---|---|---|---|---|---|---|---|
| 2.56 | 4.79 | 5.09 | 7.25 | 8.94 | 5.83 | 1.20 | 24905.00 | 0 |

When comparing the average minimum temperature to Southern California shown in Figure 34, both the counties Kern and Ventura have the lowest average minimum temperature, while Orange county has the highest minimum temperature. Comparing all three variables together, the exploratory analysis indicates Orange County has one of the highest average precipitation amounts, highest temperature maximum average, and highest temperature

minimum averages. While  Ventura County has one of the lowest average precipitation amounts, a high temperature maximum average, and one of the lowest temperature minimum averages.

**Figure 34**

*Boxplot showing temperature minimum by County*




Along with wildfire, carbon emission, and weather data, the descriptive statistics was performed on populations and WUI data as well to explore variables TotalPop, Income, Area_SquareKm, Housing_number, and Population_count.

***TotalPop And Income Variables***

Figure 35 presents the histogram of the TotalPop variable, which displays a unimodal normal distribution. The population count with the most instances, which is a center of distribution, appears to be near 4,500, and the least appears to be near 15,750. Figure 35 depicts that the variable TotalPop has a positively skewed distribution with high kurtosis, also known as leptokurtic distribution. Similarly, the histogram in Figure 36 displays a unimodal normal distribution of the variable Income from Populations dataset. The center of the histogram that represents the value of income with the greatest number of observations, appears to be near

$65,000, and $196,000 shows the least number of observations. The distribution of variable Income is positively skewed with low kurtosis, known as platykurtic distribution.

**Figure 35**

*Distribution of TotalPop*

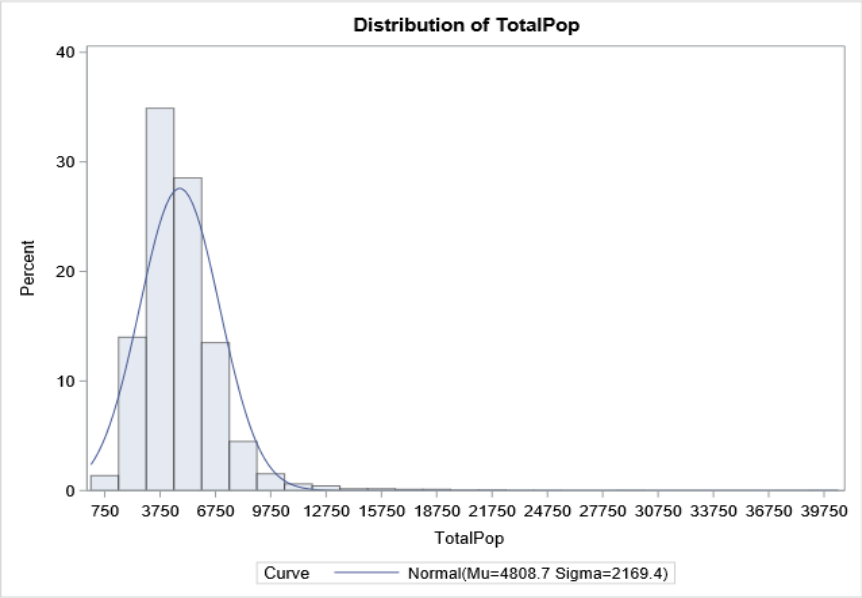
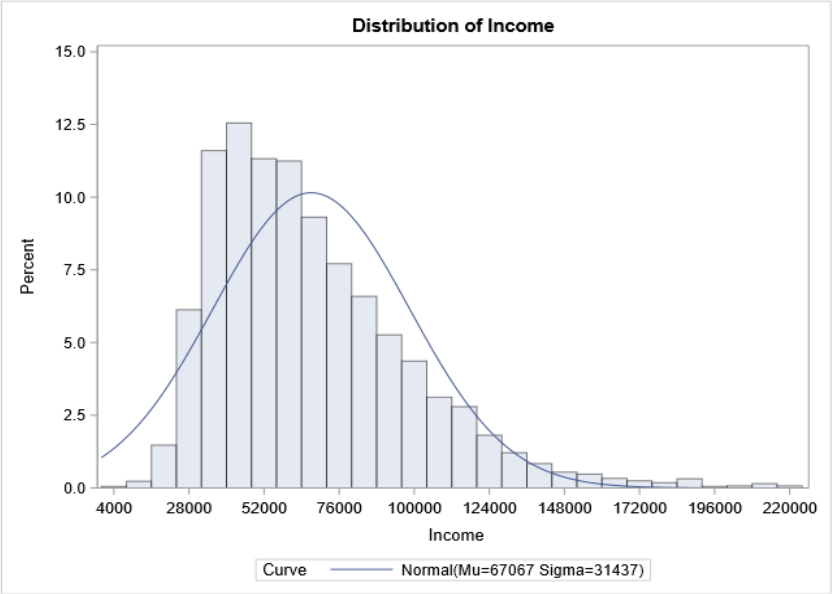



**Figure 36**

*Distribution of Income*



Descriptive statistics were calculated on 6,120 observations of the Population dataset using the Summary Statistics task in SAS Enterprise Guide (SAS EG), and the output is shown in table 9. The mean of total population (TotalPop) for all observations is 4,808.73, whereas the mean of Income is 67,067.41. The Standard Error for variables TotalPop and Income are 27.73 and 401.84 respectively.

**Table 9**

*Summary statistics of variables TotalPop and Income*

| Variable | Mean | Std Dev | Min | Max | N | N Miss | Std Error | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| TotalPop | 4808.73 | 2169.41 | 6 | 40402 | 6120 | 0 | 27.73 | 4706319.24 | 2.87 | 23.42 |
| Income | 67067.41 | 31436.57 | 5645 | 221635 | 6120 | 0 | 401.84 | 988257837 | 1.1869718 | 1.93 |

County has been used as a classification variable while generating a summary statistics table that helps in understanding the distribution of the data in variables TotalPop and Income for each county in Southern California. Tables 10 and 11 provide these statistics for both the variables respectively. The Populations dataset contains 612 observations per county without any missing value in variables TotalPop and Income.

**Table 10**

*Summary statistics of TotalPop using County as classification variable.*

| County | N Obs | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness |
|---|---|---|---|---|---|---|---|---|---|
| Imperial | 612 | 4384.93 | 1552.81 | 643 | 12986 | 612 | 0 | 2411208.46 | 1.05 |
| Kern | 612 | 4841.4 | 2051.13 | 58 | 18533 | 612 | 0 | 4207140.40 | 2.19 |
| Los Angeles | 612 | 3974.28 | 1204.93 | 1065 | 8229 | 612 | 0 | 1451852.40 | 0.64 |
| Orange | 612 | 5489.6 | 2526.68 | 27 | 24036 | 612 | 0 | 6384101.78 | 2.38 |
| Riverside | 612 | 5202.81 | 2479.62 | 151 | 17915 | 612 | 0 | 6148515.07 | 1.35 |
| San Bernardino | 612 | 5763.13 | 2555.91 | 1044 | 23877 | 612 | 0 | 6532650.57 | 2.42 |
| San Diego | 612 | 5223.17 | 2880.15 | 23 | 40402 | 612 | 0 | 8295251.66 | 5.14 |
| San Luis Obispo | 612 | 4376.32 | 1776.01 | 200 | 16028 | 612 | 0 | 3154227.30 | 1.68 |
| Santa Barbara | 612 | 4141.84 | 1489.44 | 6 | 12917 | 612 | 0 | 2218420.50 | 1.25 |
| Ventura | 612 | 4689.85 | 1760.54 | 31 | 12653 | 612 | 0 | 3099507.28 | 0.22 |

**Table 11**

*Summary statistics of Income using County as classification variable*

| County | N Obs | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness |
|---|---|---|---|---|---|---|---|---|---|
| Imperial | 612 | 63594.75 | 35418.63 | 5682 | 221635 | 612 | 0 | 1254479419 | 1.57 |
| Kern | 612 | 60882.53 | 26543.2 | 21746 | 205170 | 612 | 0 | 704541262 | 1.43 |
| Los Angeles | 612 | 61555.83 | 29654.26 | 10262 | 187891 | 612 | 0 | 879375328 | 1.45 |
| Orange | 612 | 85757.32 | 32320.85 | 24211 | 209095 | 612 | 0 | 1044637638 | 0.71 |
| Riverside | 612 | 64506.93 | 24946.5 | 20394 | 149091 | 612 | 0 | 622327833 | 0.68 |
| San Bernardino | 612 | 61051.77 | 24786.65 | 14550 | 139805 | 612 | 0 | 614378223 | 0.86 |
| San Diego | 612 | 75406.9 | 30383.96 | 22614 | 183929 | 612 | 0 | 923185288 | 0.76 |
| San Luis Obispo | 612 | 67327.72 | 29856.02 | 5645 | 191642 | 612 | 0 | 891382190 | 1.29 |
| Santa Barbara | 612 | 59117.13 | 36262.8 | 5682 | 221635 | 612 | 0 | 1314990838 | 1.73 |
| Ventura | 612 | 71473.21 | 31966.16 | 16196 | 207679 | 612 | 0 | 1021835105 | 1.10 |

The boxplot represented by Figure 37 has been generated using PROC SGPLOT in SAS studio and shows the distribution of quantitative data in a way that facilitates comparisons between variables TotalPop and County_Code; each box represents populations count for the specific county. Similarly, the boxplot shown in Figure 38 gives an idea about variation of Income based on County_Code in southern California. The boxplots in both the figures 37 and 38 depict that a few data points are located outside the whiskers of boxplots for both the variables TotalPop and Income, i.e., the variables consist of outliers. However, the outliers have not been excluded or modified because the data points belong to census tracts which are small statistical subdivisions of each county. These extreme data points have added value in the prediction model to determine whether variation in population count and income affect the frequency and severity of wildfires in that county.

**Figure 37**

*Box plot of variables TotalPop and County*

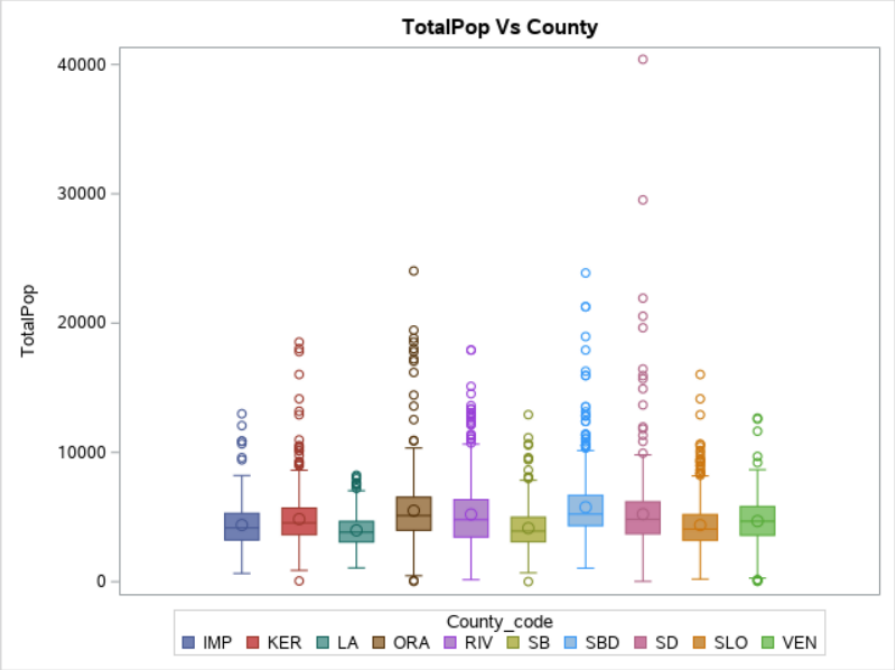
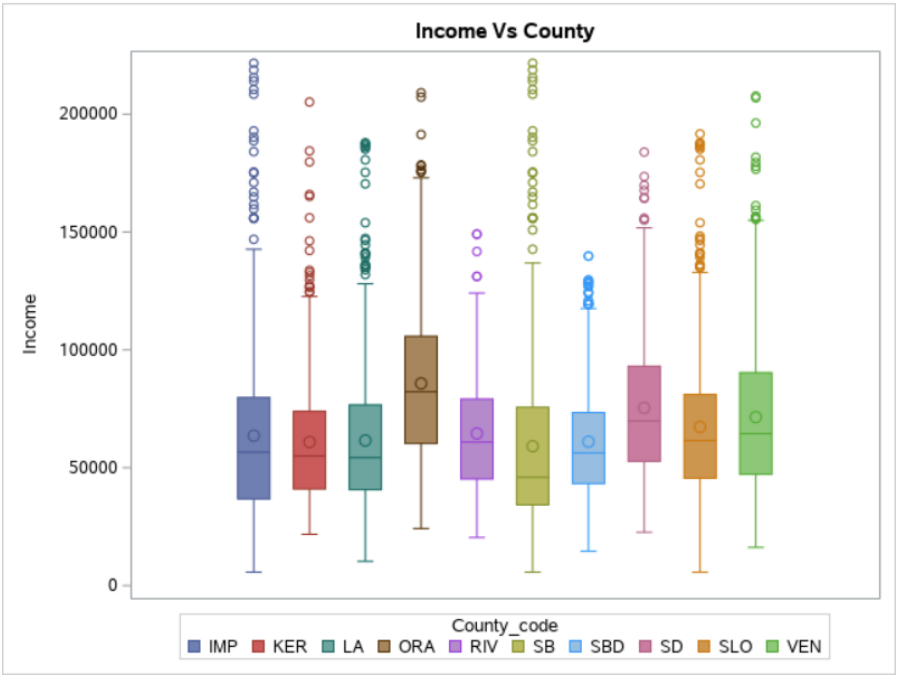



**Figure 38**

*Box plot of variables Income and County*



***Area_SquareKm, Housing_number, and Population_count***

Table 12 shows that the value of Area_SquareKm for a large number of observations appears to be more than yet close to 600. Histograms created using SAS EG depict distribution of variables Area_SquareKm, Housing_number, and Population_count from the WUI dataset are shown in Figure 39, Figure 40, and Figure 41, and all of them display positively skewed unimodal normal distribution. Although Figures 40 and 41 present that Housing_number and Population_count variables contain the greatest number of observations with or near value 0, table 12 clearly depicts that the mean values of both are 12,127.44 and 28,953.81 respectively.

**Table 12**

*Summary statistics of Area_SquareKm, Housing_number, and Population_count*

| Variable | Mean | Std Dev | Min | Max | N |
|---|---|---|---|---|---|
| Area_SquareKm | 667.02 | 1994.20 | 0.00 | 50328.90 | 24960 |
| Housing_number | 12127.44 | 52047.38 | 0.00 | 2884150.00 | 24960 |
| Population_count | 28953.81 | 136194.24 | 0.00 | 8278921.00 | 24960 |

**Figure 39**

*Distribution of Area_SquareKm*

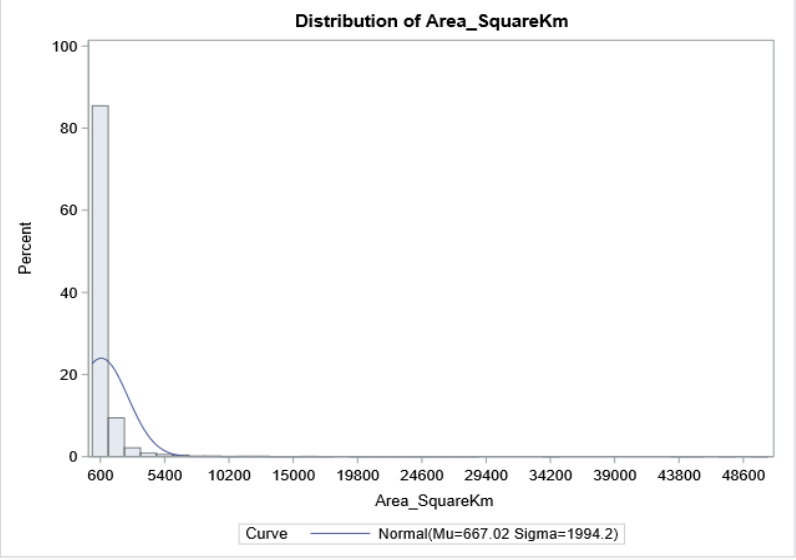


**Figure 40**

*Distribution of Housing_number*

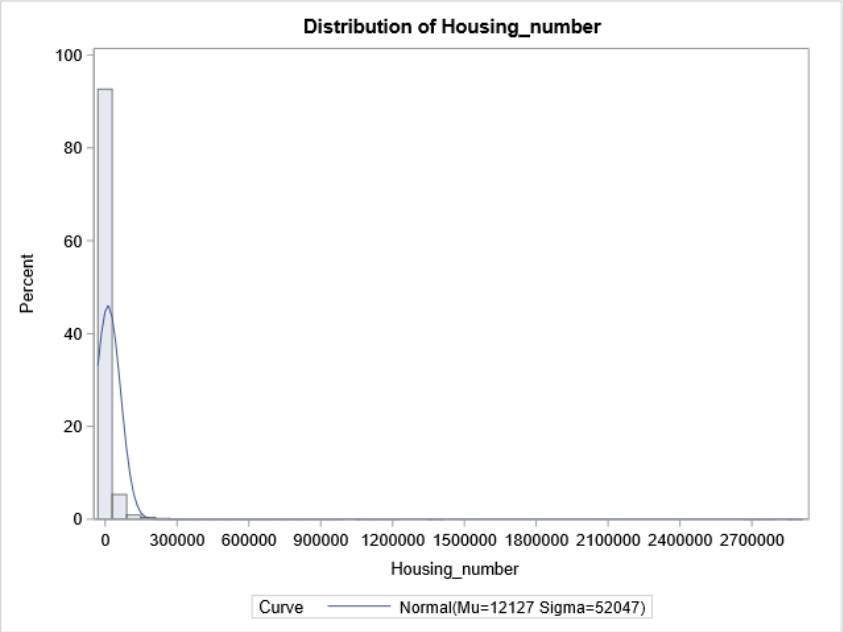
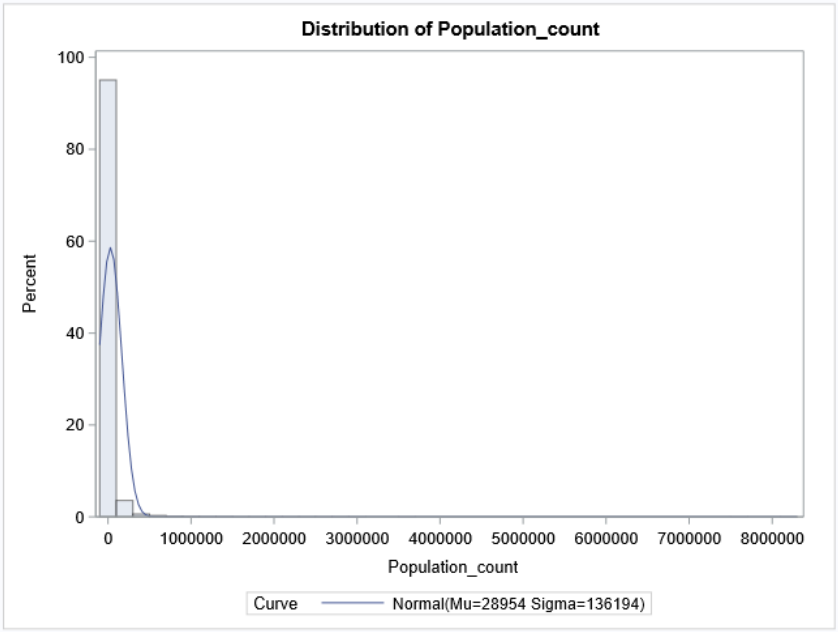



**Figure 41**

*Distribution of Population_count*



Statistics such as number of observations, number of missing observations, mean, standard deviation, minimum and maximum values, and variance were calculated for variables Area_SquareKm, Housing_number, and Population_count using WUI_Category as a classification variable. These statistics are shown in tables 13, 14, and 15. Each category value of the WUI_Category variable contains 6,240 numbers of observations without any missing value.

**Table 13**

*Summary statistics of Area_SquareKm using WUI_Category as classification variable*

| WUI_Category | N Obs | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness |
|---|---|---|---|---|---|---|---|---|---|
| Interface | 6240 | 47.35 | 64.72 | 0.00 | 835.80 | 6240 | 0 | 4188.55 | 3.34 |
| Intermix | 6240 | 174.47 | 213.12 | 0.00 | 1476.20 | 6240 | 0 | 45420.57 | 1.84 |
| NonWUI | 6240 | 2224.43 | 3541.24 | 2.70 | 50328.90 | 6240 | 0 | 12540378.93 | 6.74 |
| WUI | 6240 | 221.82 | 261.64 | 0.00 | 2035.50 | 6240 | 0 | 68457.30 | 1.86 |

**Table 14**

*Summary statistics of Housing_number using WUI_Category as classification variable*

| WUI_Category | N Obs | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness |
|---|---|---|---|---|---|---|---|---|---|
| Interface | 6240 | 7347.24 | 21524.74 | 0 | 521425 | 6240 | 0 | 463314412.00 | 11.04 |
| Intermix | 6240 | 4731.04 | 7101.15 | 0 | 75829 | 6240 | 0 | 50426394.68 | 3.06 |
| NonWUI | 6240 | 24353.19 | 96943.24 | 0 | 2884150 | 6240 | 0 | 9397991202.00 | 14.80 |
| WUI | 6240 | 12078.28 | 26423.96 | 0 | 560926 | 6240 | 0 | 698225764.00 | 8.51 |

**Table 15**

*Summary statistics of Population_count using WUI_Category as classification variable*

| WUI_Category | N Obs | Mean | Std Dev | Min | Max | N | N Miss | Variance | Skewness |
|---|---|---|---|---|---|---|---|---|---|
| Interface | 6240 | 17237.64 | 54698.08 | 0 | 1431895 | 6240 | 0 | 2991879495.00 | 12.21 |
| Intermix | 6240 | 10860.25 | 16987.21 | 0 | 164346 | 6240 | 0 | 288565219.00 | 3.17 |
| NonWUI | 6240 | 59619.45 | 255243.51 | 3 | 8278921 | 6240 | 0 | 65149248190.00 | 17.68 |
| WUI | 6240 | 28097.89 | 66096.30 | 0 | 1539684 | 6240 | 0 | 4368720888.00 | 9.52 |

**Exploratory Data Analysis on Master Data**

All the cleaned datasets for components wildfire, carbon emission, weather, and population were imported into traditional tables, so that an extract, transform, and load (ETL)

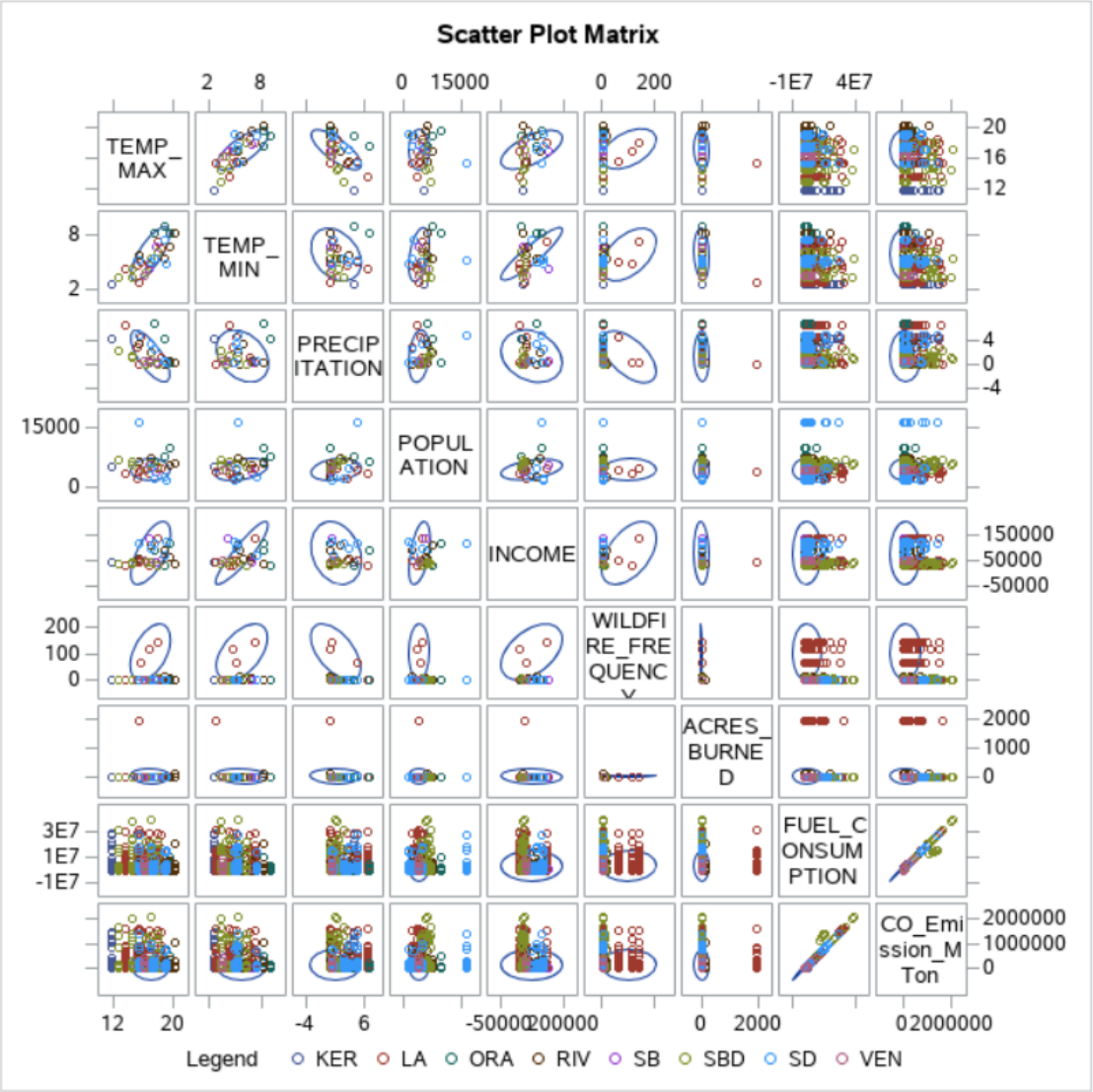
could be performed. Several queries were created to join the traditional tables and extract the data from required columns, so that it can be inserted into the fact and dimension tables of the data mart. The fact and dimension tables were then joined using common key attributes and the selected data was exported to a CSV file to create the master data file that contains data for analysis and creating the prediction model.

***Initial exploratory analysis of selected outcome and explanatory variables***

Figure 42 shows a scatter-plot matrix of all continuous variables in master data: Temp_Max, Temp_Min, Precipitation, Total_Population, Income, Wildfire_Frequency, Acres_Burned, Fuel_Consumption, and CO2_Emissions_Mettrictonnes. Each scatter plot in the matrix has been grouped by County_Name, so that each paired observation has a county-associated label. The scatter-plot matrix visually presented that precipitation is negatively correlated with maximum and minimum temperature; populations and income were slightly and positively correlated with each other; carbon emission showed a strong positive correlation with fuel consumption; finally, wildfire frequency and severity did not show any correlation with each other, however, they presented correlation with the selected explanatory variables. Descriptive statistics were generated using the Simple Statistics task in SAS EG for all outcome and explanatory variables that is presented by table 16.

**Figure 42**

*Correlation matrix of all continuous variables*




*Note*. Correlation matrix has been generated using the PROC SGSCATTER procedure in SAS studio.

**Table 16**

*Descriptive statistics of outcome and explanatory variables*

| Variable | Mean | Std Dev | Min | Max | N | Std Error | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| Wildfire_Frequency | 104.82 | 43.67 | 1 | 143.00 | 24838 | 0.28 | 1907.14 | -1.11 | 0.13 |
| Incident_Acres_Burned | 5.45 | 101.15 | 0 | 1952.00 | 24905 | 0.64 | 10230.40 | 19.18 | 366.14 |
| Temp_Max | 17.00 | 1.07 | 11.70 | 20.21 | 24905 | 0.01 | 1.14 | -0.73 | 1.46 |
| Precipitation | 1.18 | 1.75 | 0.01 | 6.79 | 24905 | 0.01 | 3.05 | 1.45 | 0.22 |
| Total_Population | 4337.71 | 1129.48 | 1600 | 16449.00 | 24905 | 7.16 | 1275729.40 | 4.88 | 52.68 |
| Co2_Emissions | 119407.78 | 256323.08 | 0 | 2059093.98 | 24905 | 1624.22 | 65701519673 | 3.41 | 12.81 |

**Outcome Variable - Wildfire_Frequency.** Correlation analysis was conducted using PROC CORR procedure in SAS studio for Wildfire_Frequency as an analysis variable and Temp_Max, Precipitation, Total_Population, and CO2_Emissions_Mettrictonnes as correlated variables. In the result of the CORR procedure shown in table 17(b), Pearson correlation coefficients confirm that variables Temp_Max, Total_Population, and CO2_Emissions_Mettrictonnes have positive correlation, whereas Precipitation has a negative correlation with Wildfire_Frequency. The *P-value* < .0001 confirms the statistical significance of correlation between Wildfire_Frequency and all the selected explanatory variables. Because the correlation coefficients of variables Temp_Max, Total_Population, and CO2_Emissions_Mettrictonnes are less than 0.5, they have lower correlation with Wildfire_Frequency. On the other hand, Precipitation has higher correlation with Wildfire_frequency, because the absolute value of correlation coefficient is 0.57512 that is greater than 0.5. The same results can be visualized by scatter plots in Figure 43(a), (b), (c), and (d).

**Table 17**

*Pearson Correlation Coefficients of Wildfire_Frequency with selected explanatory variables*

**(a)**

| The CORR Procedure | |
|---|---|
| 4 With Variables: | Temp_Max Precipitation Total_Population Co2_Emissions_Metrictonnes |
| 1 Variables: | Wildfire_Frequency |

**(b)**

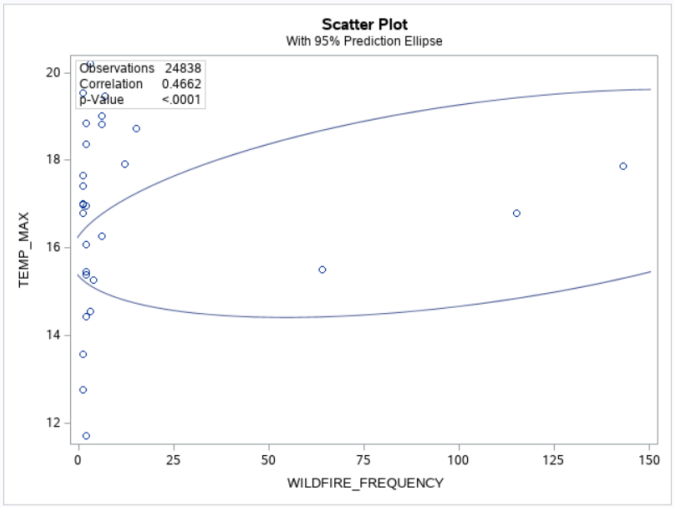
| Pearson Correlation Coefficients<br>Prob > \|r\| under H0: Rho=0<br>Number of Observations | |
|---|---|
| | WILDFIRE_FREQUENCY |
| TEMP_MAX | 0.4662 |
| | <.0001 |
| | 24838 |
| PRECIPITATION | -0.57512 |
| | <.0001 |
| | 24838 |
| TOTAL_POPULATION | 0.0371 |
| | <.0001 |
| | 24838 |
| CO2_EMISSIONS_METRICTONNES | 0.02804 |
| | <.0001 |
| | 24838 |

*Note.* (a) outcome and with variable list (b) correlation coefficients of each variable

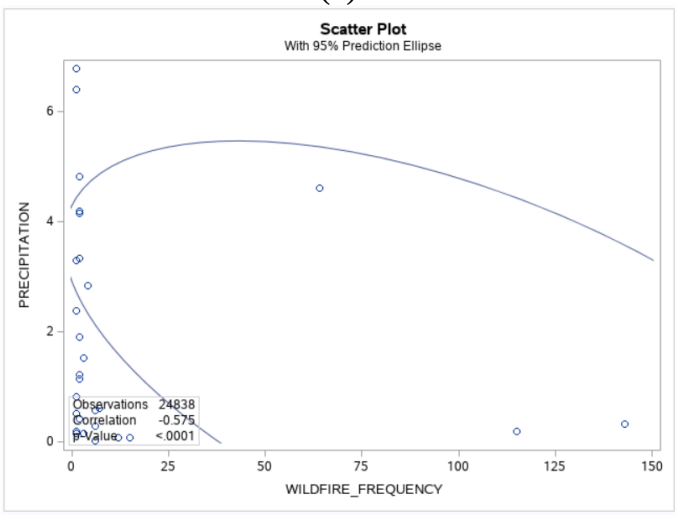
**Figure 43**

*Scatter plots of outcome variable Wildfire_Frequency with all explanatory variables*

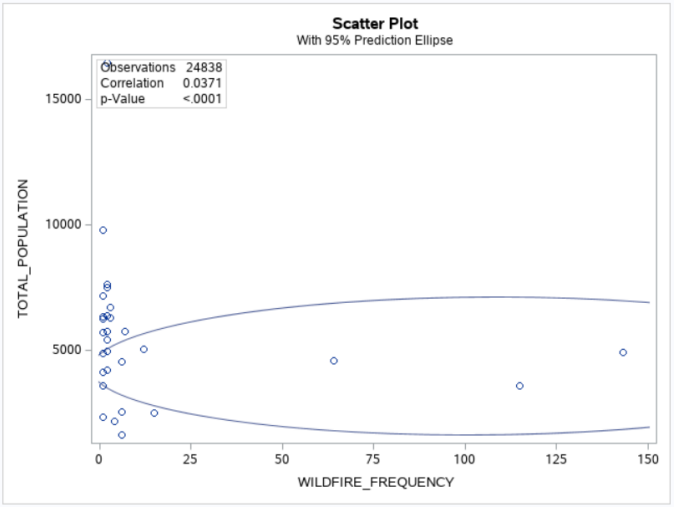
(a)



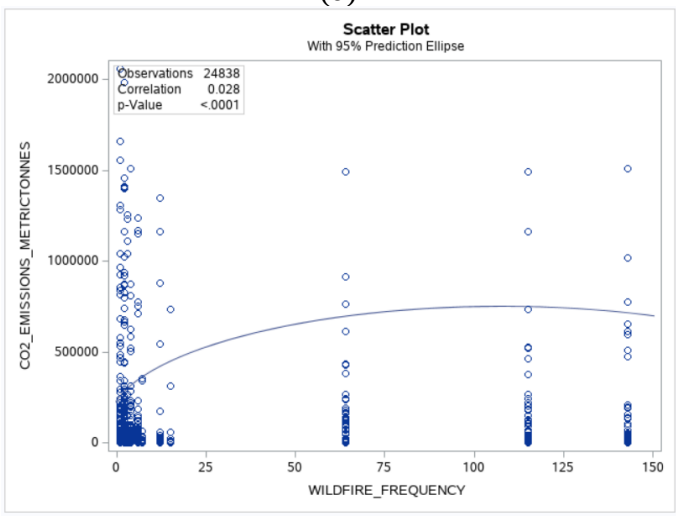

(b)




(c)




(d)




*Note:* (a) Scatter plot presenting the correlation between Wildfire_Frequency and Temp_Max. (b) Scatter plot presenting the correlation between Wildfire_Frequency and Precipitation. (c) Scatter plot presenting the correlation between Wildfire_Frequency and Total_Population. (d) Scatter plot presenting the correlation between Wildfire_Frequency and CO2_Emissions_Metrictonnes.

Table 18(a) and (b) present the results of multiple linear regression analysis that was conducted using PROC REG in SAS studio to explain Wildfire_Frequency (outcome variable) using Temp_Max, Precipitation, Total_Population, and CO2_Emissions_Mettrictonnes (explanatory variables). Table 18(a) shows that the number of observations used in the multilinear regression model is 24,838; the effect size F(4, 24833) is equal to 3,468.94, and has *P-value* < 0.0001. Therefore, the relationship between all the explanatory variables and the outcome variable is statistically significant. The value of an adjusted R-square presents that 35.84% of variance in Wildfire_Frequency is explained by all the selected explanatory variables.

Table 18(b) presents the table of parameter estimates that provides t-values of variables Temp_Max, Precipitation, Total_Population, and CO2_Emissions_MetricTonnes, and the values are 3.09, -73.32, 30.53, and 3.51 respectively. Precipitation is the strongest predictor of Wildfire_Frequency because it has the highest absolute t-value, and negative sign explains the negative correlation between Precipitation and Wildfire_Frequency. Furthermore, the analysis states that the selected explanatory variables are related to the outcome variable, and the relationship is statistically significant, because *p-value* (Pr>|t|) is less than 0.05 for all of them.

**Table 18**

*Multilinear regression results of Wildfire_Frequency with explanatory variables*

**(a)**

Model: MODEL1
Dependent Variable: WILDFIRE_FREQUENCY
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F* |
|---|---|---|---|---|---|
| Model | 4 | 16979646 | 4244911 | 3468.94 | <.0001 |
| Error | 24833 | 30387968 | 1223.693 | | |
| Corrected Total | 24837 | 47367613 | | | |

| | |
|---|---|
| Root MSE | 34.98132 |
| Dependent Mean | 104.81714 |
| Coeff Var | 33.37367 |
| R-Square | 0.3585 |
| Adj R-Sq | 0.3584 |

*Note. *p < 0.05*; Adj R-Sq = 0.3584 can be interpreted as 35.84% of variance in Wildfire_Frequency is explained by selected explanatory variables.

**(b)**

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS | Standardized Estimate |
|---|---|---|---|---|---|---|---|---|
| Intercept | 1 | 77.38787 | 5.51991 | 14.02 | <.0001 | 272885961 | 240522 | 0 |
| Temp_Max | 1 | 0.99798 | 0.32303 | 3.09 | 0.002 | 10295136 | 11680 | 0.02436 |
| Precipitation | 1 | -14.78373 | 0.20163 | -73.32 | <.0001 | 5522577 | 6578799 | -0.59207 |
| Total_Population | 1 | 0.00634 | 0.0002078 | 30.53 | <.0001 | 1146825 | 1140453 | 0.16428 |
| Co2_Emissions_Metrictonnes | 1 | 0.00000305 | 8.67E-07 | 3.51 | 0.0004 | 15107 | 15107 | 0.01787 |

*Note.* (a) Analysis of Variance table from the result of multiple linear regression analysis with Wildfire_Frequency as outcome variable. (b) Parameter Estimates table from table from the result of multiple linear regression analysis with Wildfire_Frequency as outcome variable.

**Outcome Variable – Incident_Acres_Burned.** Correlation analysis was conducted using PROC CORR procedure in SAS studio for Incident_Acres_Burned as an analysis variable

and Temp_Max, Precipitation, Total_Population, and CO2_Emissions_Mettrictonnes as correlated variables. The Pearson Correlation Coefficients are shown in table 19(b). The correlation coefficients for variables Temp_Max, Precipitation, and Total_Population are -0.07919, -0.03416, and -0.01794 respectively; the negative values state that the Incident_Acres_Burned has negative correlation with the three variables. Because *p-value* is less than 0.05 for these three explanatory variables, their correlation with Incident_Acres_Burned is statistically significant. On the other hand, variable CO2_Emissions_Mettrictonnes is positively correlated with Incident_Acres_Burned, because the correlation coefficient is 0.00339 i.e., a positive value. However, the *P-value is 0.5931 that is greater than 0.05*, hence the correlation between Incident_Acres_Burned and CO2_Emissions_MetricTonnes is not statistically significant. The same results can be visualized by scatter plots in Figure 44(a), (b), (c), and (d).

**Table 19**

*Pearson Correlation Coefficients of Incident_Acres_Burned with selected explanatory variables*

**(a)**

| | |
|---|---|
| 4 With Variables: | Temp_Max Precipitation Total_Population Co2_Emissions_Metrictonnes |
| 1 Variables: | Incident_Acres_Burned |

**(b)**

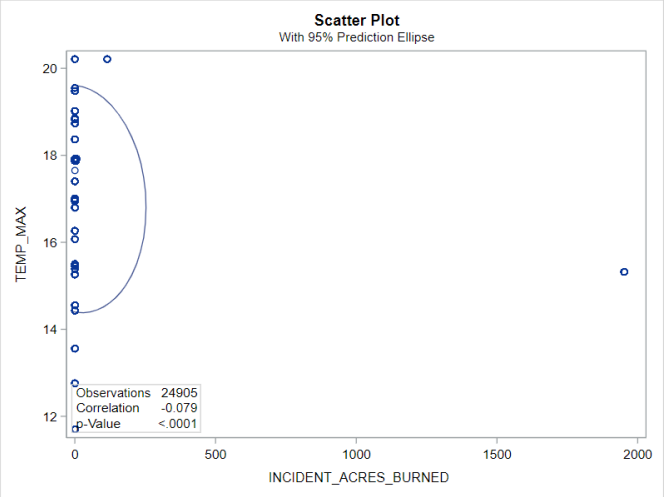
| Pearson Correlation Coefficients, N = 24905<br>Prob > \|r\| under H0: Rho=0 | |
|---|---|
| | INCIDENT_ACRES_BURNED |
| Temp_Max | -0.07919 |
| | <.0001* |
| Precipitation | -0.03416 |
| | <.0001* |
| Total_Population | -0.01794 |
| | 0.0046* |
| Co2_Emissions_Metrictonnes | 0.00339 |
| | 0.5931* |

*Note: * = p-value*

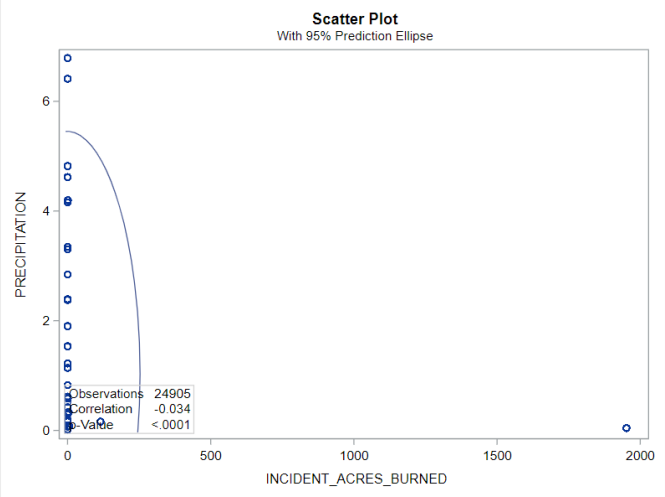
**Figure 44**

*Scatter plots of outcome variable Incident_Acres_Burned with all explanatory variables*

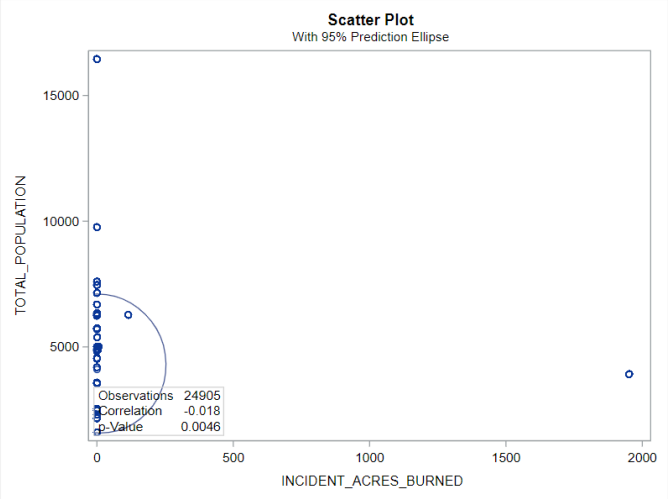
(a)



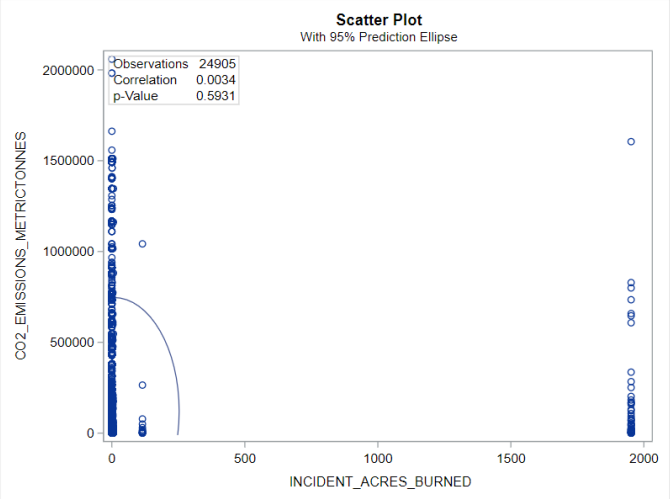

(b)




(c)




(d)




*Note.* (a) Scatter plot presenting the correlation between Incident_Acres_Burned and Temp_Max. (b) Scatter plot presenting the correlation between Incident_Acres_Burned and Precipitation. (c) Scatter plot presenting the correlation between Incident_Acres_Burned and Total_Population. (d) Scatter plot presenting the correlation between Incident_Acres_Burned and CO2_Emissions_Metrictonnes.

Table 20(a) and (b) present the results of multiple linear regression analysis that was conducted using PROC REG in SAS studio to explain Incident_Acres_Burned (outcome variable) using Temp_Max, Precipitation, Total_Population, and CO2_Emissions_Mettrictonnes (explanatory variables). Table 20(a) shows that the number of observations used in the multilinear regression model is 24,905; the effect size F(4, 24900) is equal to 166.96, and has *P-value* < 0.0001. Therefore, Incident_Acres_Burned has a statistically significant relationship with the selected explanatory variables: Temp_Max, Precipitation, Total_Population, and Co2_Emissions_Metrictonnes. The value of an adjusted R-square presents that 2.60% of variance in Incident_Acres_Burned is explained by the selected explanatory variables.

Table 20(b) presents the table of parameter estimates that provides t-values of variables Temp_Max, Precipitation, Total_Population, and CO2_Emissions_MetricTonnes; the values are -25.19, -22.35, 4.48, and -0.37 respectively. Temp_Max is the strongest predictor of Incident_Acres_Burned because it has the highest absolute t-value. The negative sign of t-values for Temp_Max and Precipitation explains their negative correlation with the outcome variable. Furthermore, the analysis states that the Temp_Max, Precipitation, and Total_Population are related to the outcome variable, and the relationship is statistically significant, because *p-value* (Pr>|t|) is less than 0.05 for them. In contrast, CO2_Emissions_MetricTonnes does not have a statistically significant relationship with Incident_Acres_Burned due to the *p-value* (Pr > |t|) which is greater than 0.05 and 95% confidence interval crosses over zero.

**Table 20**

*Multilinear regression results of Incident_Acres_Burned with explanatory variables*

**(a)**

| Analysis of Variance (Dependent Variable: **Incident_Acres_Burned**) | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F* |
| Model | 4 | 6654847 | 1663712 | 166.96 | <.0001 |
| Error | 24900 | 248122935 | 9964.7765 | | |
| Corrected Total | 24904 | 254777782 | | | |

| | |
|---|---|
| Root MSE | 99.82373 |
| Dependent Mean | 5.45093 |
| Coeff Var | 1831.31433 |
| R-Square | 0.0261 |
| Adj R-Sq | 0.026 |

*Note.* *p < 0.05; Adj R-Sq = 0.026 can be interpreted as 2.6% of variance in Incident_Acres_Burned i.e. severity of wildfires is explained by selected explanatory variables

**(b)**

| Parameter Estimates | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS | Standardized Estimate | 95% Confidence Limits | |
| Intercept | 1 | 398.12 | 15.54 | 25.62 | <.0001 | 739994 | 6539687 | 0 | 367.66487 | 428.58698 |
| Temp_Max | 1 | -22.92 | 0.90 | -25.19 | <.0001 | 1597789 | 6325266 | -0.24207 | -24.70693 | -21.14016 |
| Precipitation | 1 | -12.72 | 0.56 | -22.35 | <.0001 | 4838242 | 4977141 | -0.21993 | -13.84525 | -11.61253 |
| Total_Population | 1 | 0.002 | 0.0005 | 4.68 | <.0001 | 217482 | 218127 | 0.03097 | 0.00161 | 0.00394 |
| CO2_Emission | 1 | -9.03E-07 | 2.47E-06 | -0.37 | 0.71 | 1333.62 | 1333.62 | -0.00229 | -5.74E-06 | 3.94E-06 |

**One-way MANOVA to determine county level differences in outcome variables at the 0.05 level.**

Outcome Variable : Wildfire_Frequency

Null Hypothesis: Mean of all the groups is the same.

$H_0 : \mu_{KER} = \mu_{LA} = \mu_{ORA} = \mu_{RIV} = \mu_{SBD} = \mu_{SD} = \mu_{SB} = \mu_{VEN}$

Alternative Hypothesis: $H_a$ : At least one of the means is different.

From the F-statistics shown in table 21(c), the null hypothesis was rejected because *p-value* < .0001 i.e., less than 0.05. Hence, variable Wildfire_Frequency has significant differences for counties in Southern California at the 0.05 level.

**Table 21**

*Analysis of variance for variable Wildfire_Frequency*

**(a)**

The GLM Procedure
Dependent Variable: INCIDENT_ACRES_BURNED

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 7 | 9083.3686 | 1297.6241 | 148.25 | <.0001 |
| Error | 24830 | 217329.8608 | 8.7527 | | |
| Corrected Total | 24837 | 226413.2294 | | | |

**(b)**

| R-Square | CoeffVar | RootMSE | Incident_Acres_Burned Mean |
|---|---|---|---|
| 0.040119 | 1478.089 | 2.958498 | 0.200157 |

**(c)**

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COUNTY_CODE | 7 | 22271482.15 | 3181640.31 | 3147.9 | <.0001 |

Outcome Variable : Incident_Acres_Burned

Null Hypothesis: Mean of all the groups is the same.

$H_0 : \mu_{KER} = \mu_{LA} = \mu_{ORA} = \mu_{RIV} = \mu_{SBD} = \mu_{SD} = \mu_{SB} = \mu_{VEN}$

Alternative Hypothesis: $H_a$ : At least one of the means is different.

From the F-statistics shown in table 22(c), the null hypothesis was rejected because *p-value* < .0001 i.e., less than 0.05. Hence, variable Incident_Acres_Burned has significant differences for counties in Southern California at the 0.05 level.

**Table 22**

*Analysis of variance for variable Incident_Acres_Burned*

**(a)**

| The GLM Procedure<br>Dependent Variable: Incident_Acres_Burned | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 7 | 9083.3686 | 1297.6241 | 148.25 | <.0001 |
| Error | 24830 | 217329.8608 | 8.7527 | | |
| Corrected Total | 24837 | 226413.2294 | | | |

**(b)**

| R-Square | Coeff Var | Root MSE | Incident_Acres_Burned Mean |
|---|---|---|---|
| 0.040119 | 1478.089 | 2.958498 | 0.200157 |

**(c)**

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| COUNTY_CODE | 7 | 9083.368609 | 1297.62409 | 148.25 | <.0001 |

**Null Hypothesis for MANOVA test criteria:** Means of all counties for both the outcome variables: Wildfire_Frequency and Incident_Acres_Burned are same.

$$H_0 = \begin{bmatrix} \mu_{KER:Wildfire\ Frequency} \\ \mu_{LA:Wildfire\ Frequency} \\ . \\ . \\ \mu_{VEN:Wildfire\ Frequency} \end{bmatrix} = \begin{bmatrix} \mu_{KER:Incident\ Acres\ Burned} \\ \mu_{LA:Incident\ Acres\ Burned} \\ . \\ . \\ \mu_{VEN:Incident\ Acres\ Burned} \end{bmatrix}$$

From table 23(c) that presents the F-approximation for all the MANOVA test criteria; null hypothesis is rejected because of *p-value* which is less than 0.05 for all the criteria.

**Table 23**

*MANOVA output tables*

**(a)**

The GLM Procedure
Multivariate Analysis of Variance
H = Type III SSCP Matrix for County_Code

| | Wildfire_Frequency | Incident_Acres_Burned |
|---|---|---|
| Wildfire_Frequency | 22271482.1 | -210740.1174 |
| Incident_Acres_Burned | -210740.117 | 9083.368609 |

**(b)**

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for County_Code
E = Error SSCP Matrix

| Characteristic Root | Percent | Characteristic Vector V'EV=1 | |
|---|---|---|---|
| | | Wildfire_Frequency | Incident_Acres_Burned |
| 0.89685495 | 96.53 | 0.00019851 | -0.00022376 |
| 0.03227771 | 3.47 | 0.00002094 | 0.00213336 |

**(c)**

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall County_Code Effect
H = Type III SSCP Matrix for County_Code
E = Error SSCP Matrix
S=2 M=2 N=12413.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.51070408 | 1416.37 | 14 | 49658 | <.0001 |
| Pillai's Trace | 0.50408 | 1195.28 | 14 | 49660 | <.0001 |
| Hotelling-Lawley Trace | 0.92913267 | 1647.77 | 14 | 39723 | <.0001 |
| Roy's Greatest Root | 0.89685495 | 3181.27 | 7 | 24830 | <.0001 |

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

## Prediction Models

### *Regression Model*

Using the Supernova package in R, the most influential variables to use in regression models were determined. Figure 64 shows what the Supernova results look like. After the variables were selected, each model was created and tested to determine effectiveness. There was

a total of four regression models created, two linear and two logistic models, because there were two outcome variables and a mix of quantitative and qualitative variables as explanatory variables.

**Table 24**

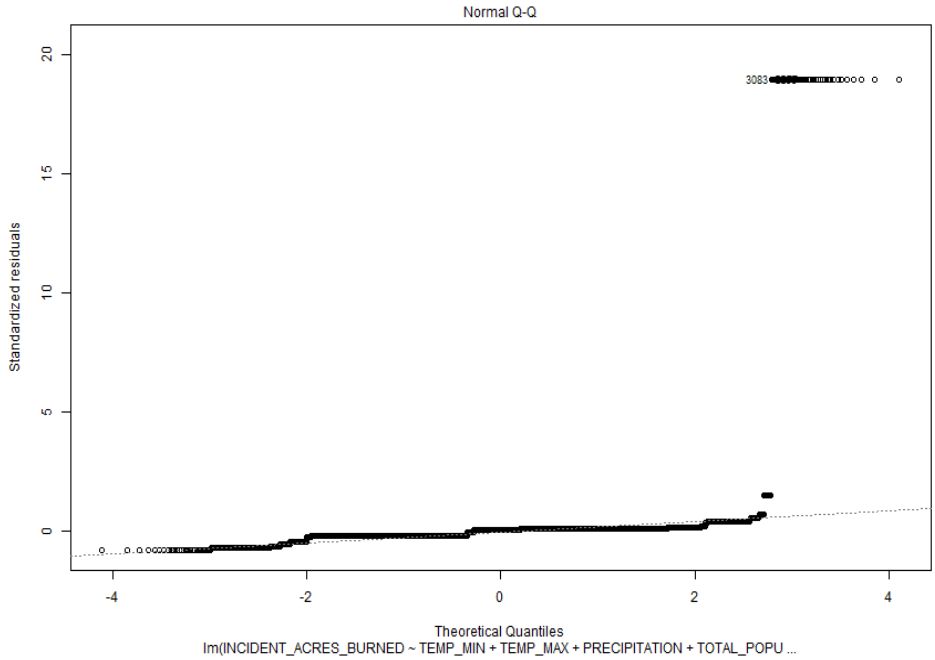
*Supernova results for Regression model example*

| | | SS | df | MS | F | PRE | P |
|---|---|---|---|---|---|---|---|
| Model | (error reduced) | 7568203.36 | 4 | 1892050.84 | 190.57 | 0.0297 | 0.0000 |
| TEMP_MIN | \| | 914689.67 | 1 | 914689.67 | 92.13 | 0.0037 | 0.0000 |
| TEMP_MAX | \| | 414791.47 | 1 | 414791.47 | 41.78 | 0.0017 | 0.0000 |
| PRECIPITATION | \| | 2262351.61 | 1 | 2262351.61 | 227.87 | 0.0091 | 0.0000 |
| TOTAL_POPULATION | \| | 819053.48 | 1 | 819053.48 | 82.49 | 0.0033 | 0.0000 |
| Error | (From Model) | 247209578.7 | 24900 | 9928.09 | | | |
| Total | (Empty Model) | 254777782.1 | 24904 | | | | |

**Linear Regression with Outcome Variable INCIDENT_ACRES_BURNED.** The first linear regression model used INCIDENT_ACRES_BURNED as the outcome variable to display the severity of the wildfires. The explanatory variables were TEMP_MIN, TEMP_MAX, PRECIPITATION, and TOTAL_POPULATION. Carbon emissions were excluded because the P values were above 0.5, meaning they were not significant. The results of the model were displayed in Figure 45. The model used Linear Model (lm) in the code references in the theoretical quantities of Figure 45, and all variables are qualitative. When analyzing the model, there was a slight positive trend between the outcome variable and the explanatory variables. So, the data indicates that there was a slight correlation between the severity of wildfires and temperature, precipitation, and population. The linear regression model of outcome variable and explanatory variables did not have a strong correlation; therefore, there was little to be predicted from this model. Therefore, weather and people have a slight effect on the severity of wildfires in Southern California.

**Figure 45**

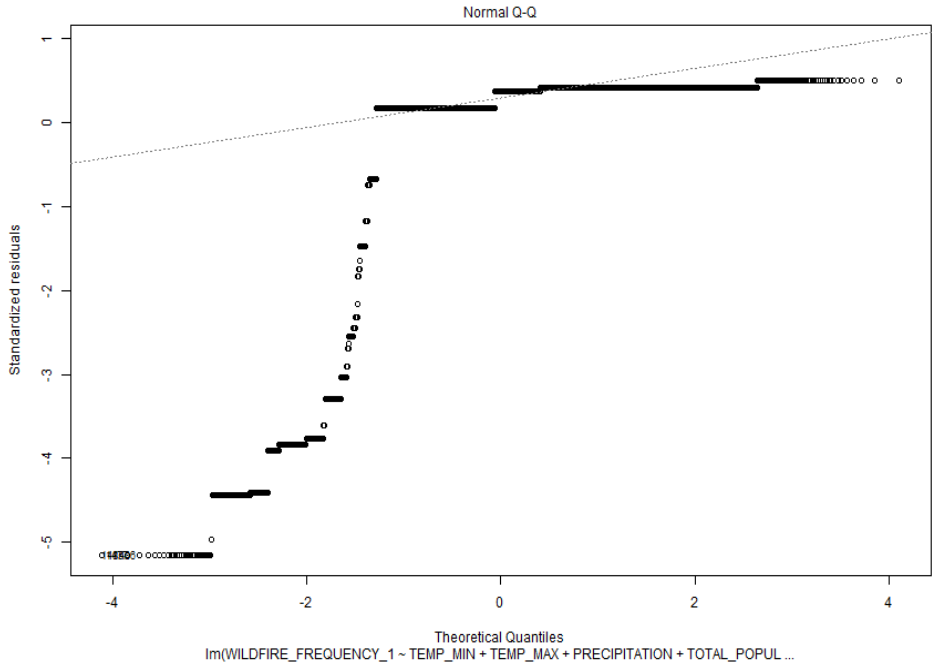
*Multiple Linear Regression Model of INCIDENT_ACRES_BURNED*




**Linear Regression with Outcome Variable WILDFIRE_FREQUENCY.** The second linear model used WILDFIRE_FREQUENCY as the outcome variable to display the frequency of the wildfires. The explanatory variables were TEMP_MIN, TEMP_MAX, PRECIPITATION, and TOTAL_POPULATION. Figure 46 showed the results of the linear regression model. There was a much stronger correlation between the outcome variable and the explanatory variables. Using Figure 46 as a reference it was assessed that the frequency of wildfires was severely impacted by weather and people.

**Figure 46**

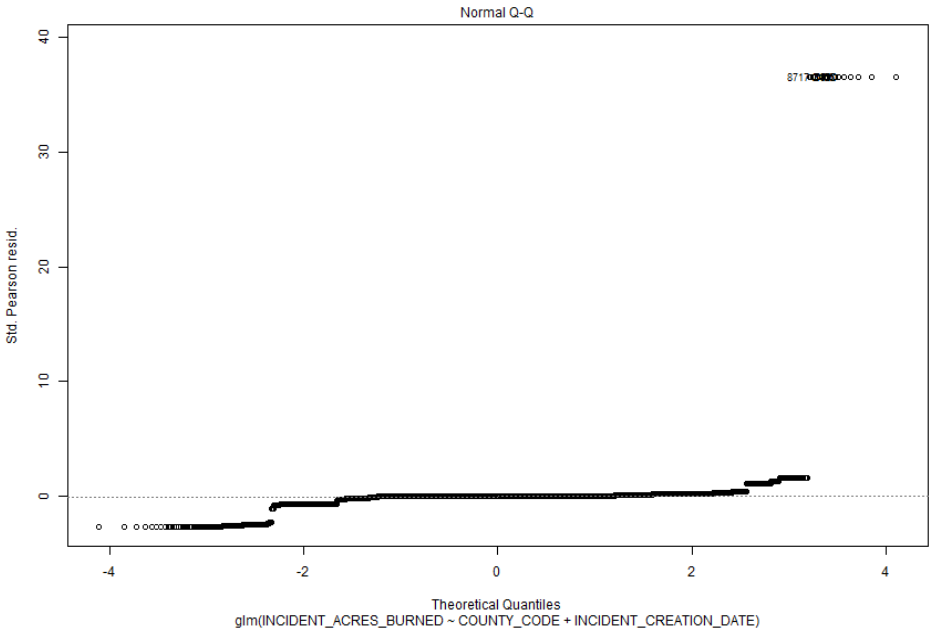
*Multiple Linear Regression Model of WILDFIRE_FREQUENCY*




**Logistic Regression with Outcome Variable : INCIDENT_ACRES_BURNED.** The first logistic regression model used INCIDENT_ACRES_BURNED as the outcome variable to display the severity of the wildfires. The explanatory variables were COUNTY_CODE and INCIDENT_CREATION_DATE. The results shown in Figure 47 indicate there was a single cutoff point where the severity of wildfires becomes affected by location or time frame. There were areas of Southern California that have more severe wildfires and times of year when wildfires were more severe. Summer months and locations with high populations tended to have more severe wildfires than low population areas during colder months.

**Figure 47**

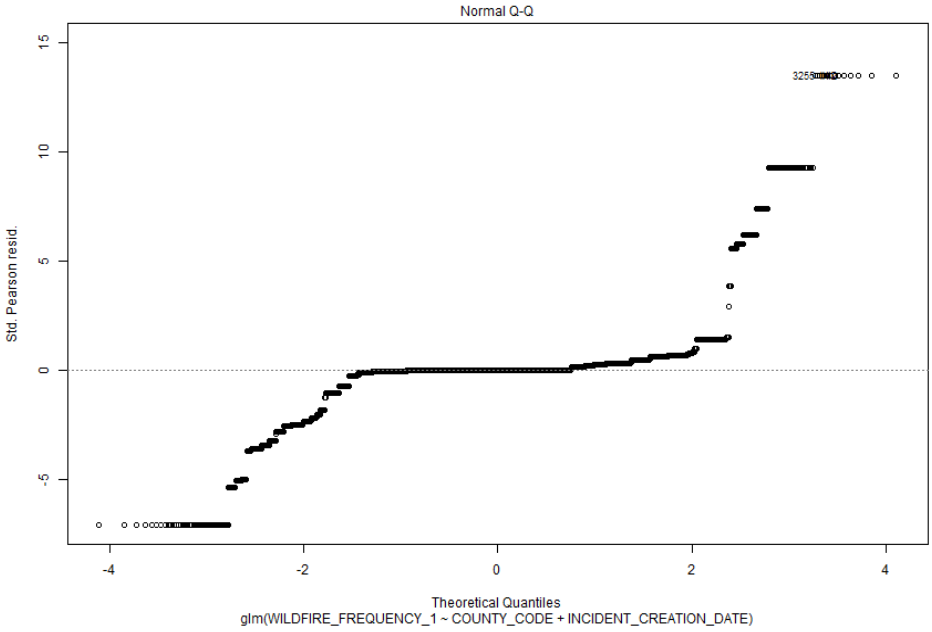
*Multiple Logistic  Regression Model of INCIDENT_ACRES_BURNED*




**Logistic Regression with outcome variable WILDFIRE_FREQUENCY.** The second logistic regression model used WILDFIRE_FREQUNCY as the outcome variable to display the frequency of the wildfires. The explanatory variables were COUNTY_CODE and INCIDENT_CREATION_DATE. The results were displayed in Figure 48 and indicated a clear delineation cutoff between affected areas. Again, There were areas of Southern California that had more frequent wildfires due to season and population. Summer months and locations with high populations tended to have more frequent wildfires than low population areas during colder months.

**Figure 48**

*Multiple Logistic  Regression Model of WILDFIRE_FREQUENCY*



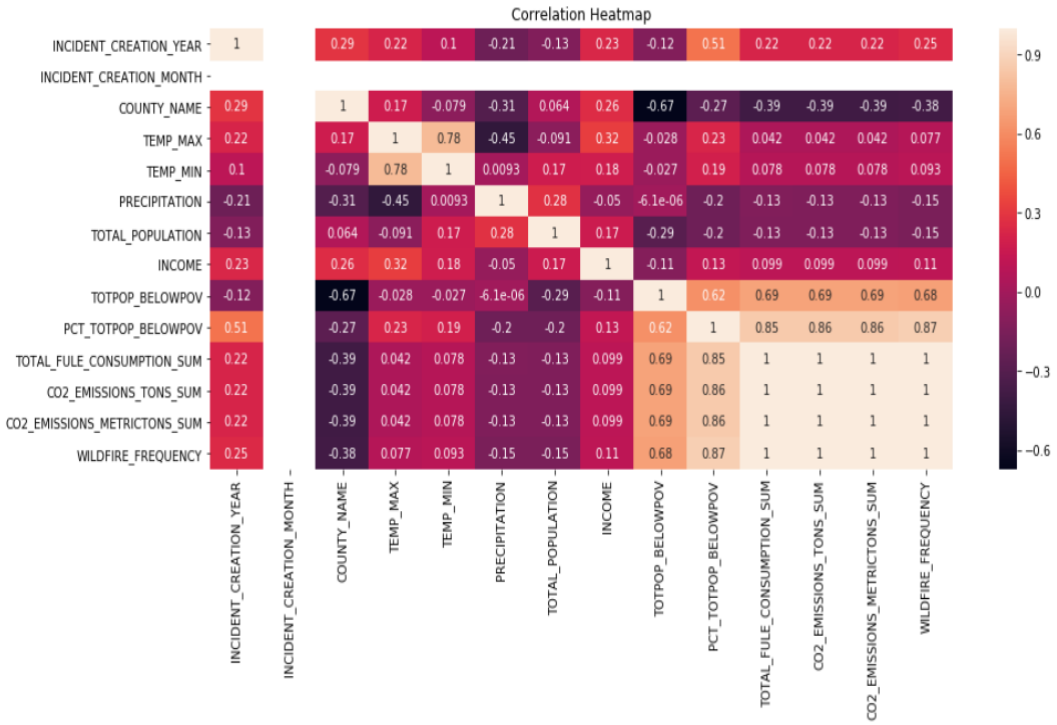

***XGBoost Prediction Model for Wildfire Frequency***

Before building the XGBoost prediction model for the wildfire frequency, all the available variables were used to create a heatmap and evaluate the existing correlation between each variable. Since XGBoost cannot process object datatype, a dummy variable was created for variables such as COUNTY_NAME. The Correlation Heatmap shown in Figure 49 displays a strong positive correlation between TOPPOP_BELOWPOV, TOTAL_FUEL_CONSUMPTION_SUM, CO2_EMISSION_TONS_SUM, CO2_EMISSION_METRICTONS_SUM, and WILDFIRE_FREQUENCY.  These findings support the hypothesis of this research that carbon emission drives wildfire frequency. Based on the correlation heatmap run on this dataset, the higher the carbon emission quantity the higher the wildfire frequency. Additionally, there is a strong negative correlation observed between Temp_Max and Precipitation, which supports the initial hypothesis that the increase in the daily

temperatures leads to lower monthly precipitation and consequently to drought in the Southern California region. Lastly, there is a negative correlation observed between Total_Population and Wildfire_Frequency, indicating that a higher population leads to lower wildfire frequency. This finding also supports the hypothesis that highly populated cities with lower WUI are less prone to wildfires. The Incident_Create_Month contains data only from January, thus, didn't produce any correlation values with other variables.

**Figure 49**

*Heatmap of all the continuous variables*

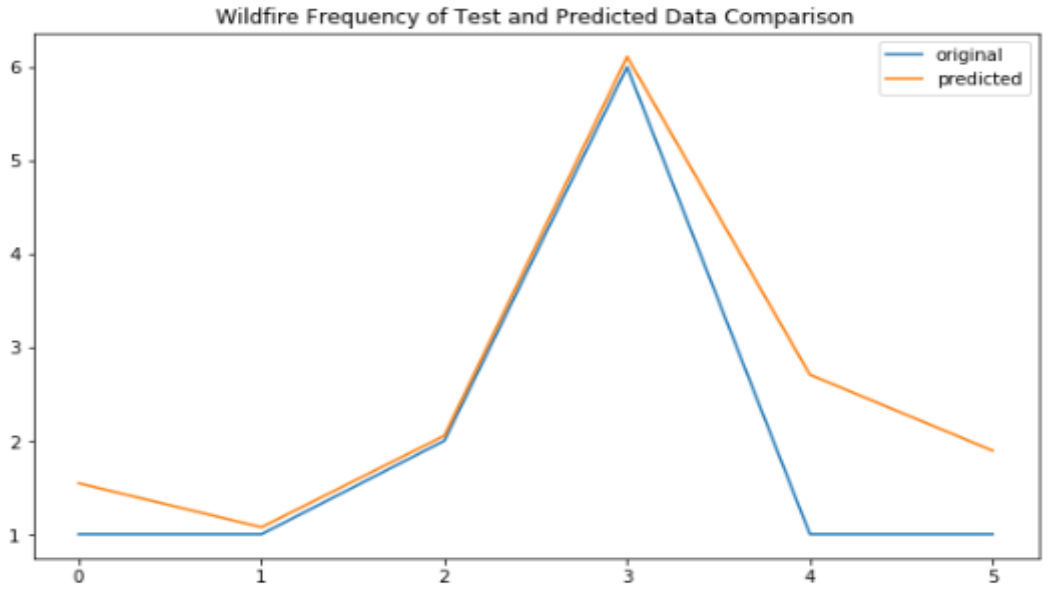



To train the XGBoost model, the dataset was divided into 80/20 training–testing sub-datasets using train_test_split function from sklearn.model_selection python package. Since the target variable (WILDFIRE_FREQUENCY) is a continuous variable, XGBRegressor from the xgboost python package was used with a linear regression objective. The model was adjusted in three different rounds to obtain the most optimal parameter settings. In the first round, most of

the parameters were kept in default settings, and the outcome model produced mean-squared-error=102.307 as shown in Table 25. In the second round, the value of gamma was modified from 1 to 0.1, and the value was increased from 1 to 1.5. This parameter setting combination improved the model by 5.4 points bringing down the mean-squared error to 96.886. In the third round, the gamma parameter was increased to 0.3 from 0.1 and colsample_bytree increased from 0.3 to 0.4. These adjustments improved the model and decreased the mean-squared error to 2.65. Lastly, in round six, the subsampling for every tree by increasing the colsample_bytree to 0.6, producing the most optimal parameter setting with mean-squared error = 0.67 and r-square = 0.798, where colsample_bytree = 0.6, learning_rate = 0.3, gamma=0.3, min_child_weight=1.5, max_depth = 6, alpha = 10, n_estimators = 13. The r-squared value of the final model indicates that about 80% of the wildfire frequency can be explained by the model. Figure 50 shows the comparison of the wildfire frequency of the test and predicted data.

**Figure 50**

*Wildfire Frequency of Test and Predicted Data Comparison*

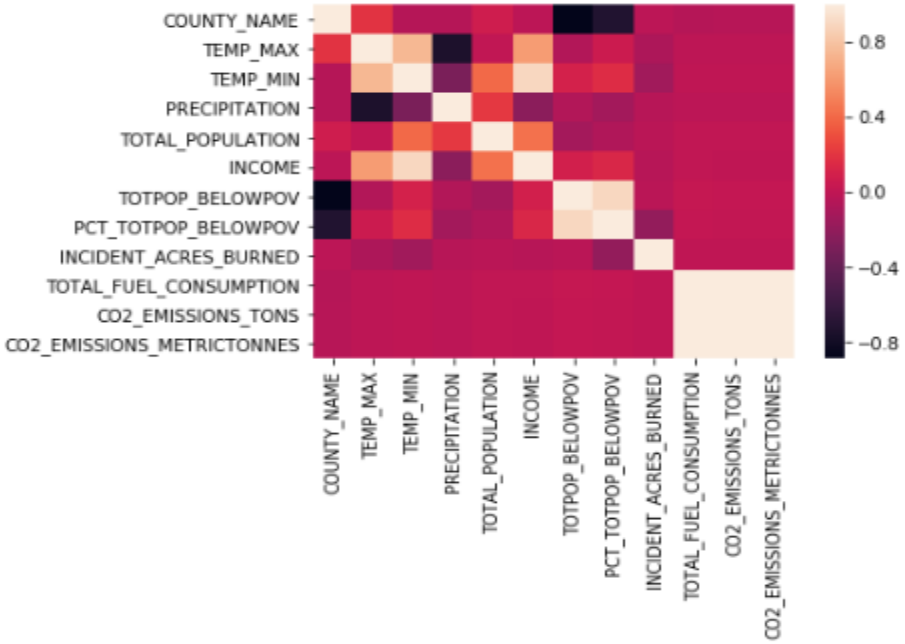


***XGBoost Prediction Model for Wildfire Severity***

All the available variables were used to create a heatmap and evaluate the existing correlation between them before building the XGBoost prediction model for the wildfire severity. The heatmap in Figure 51 displays a strong positive correlation for the variables TOTAL_FUEL_CONSUMPTION, CO2_EMISSIONS_TON, and CO2_EMISSIONS_METRICTONNES. Moreover, a strong negative correlation between the variables TEMP_MAX and PRECIPITATION was found, suggesting that a higher temperature area is associated with less rainfall, furthering more possibility of natural wildfires to occur.

**Figure 51**

*Correlation Heatmap of INCIDENT_ACRES_BURNED before model optimization*

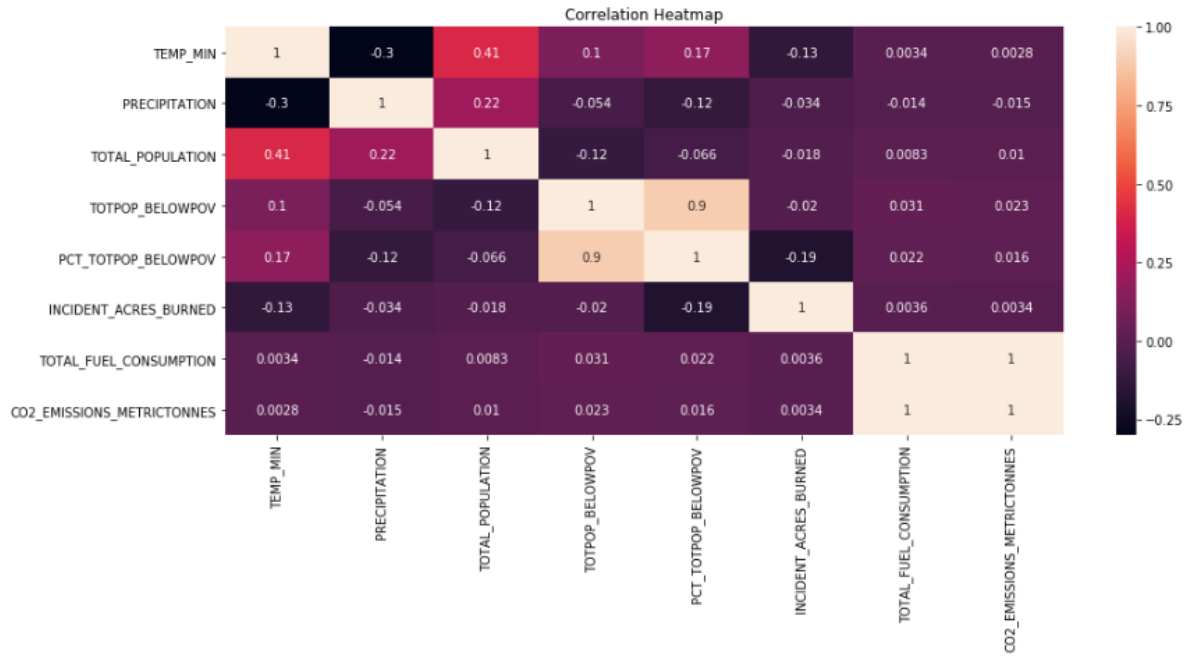



In the model optimization process, predictors TEMP_MAX, CO2_EMISSIONS_TONS, INCOME, and COUNTY_NAME were removed individually to examine the effect of each variable on the model. Figure 52 presents the correlation heatmap after the model optimization

was performed and indicates that the variable INCIDENT_ACRES_BURNED shows a strong negative correlation with variables TEMP_MIN and PCT_TOTPOP_BELOWPOV; similarly, it shows a weak negative correlation with TOTAL_POPULATION and PRECIPITATION. These results support the hypothesis of this research that minimum temperature, precipitation and population below poverty level do not contribute to the increasing acres burned by wildfires in Southern California. On the other hand, TOTAL_FUEL_CONSUMPTION, CO2_EMISSION_METRICTONS are positively correlated with INCIDENT_ACRES_BURNED. Although the positive correlation between acres burned and CO2 emission variables appears to be weak, it supports the research hypothesis that severity of wildfire is explained by carbon emission in the environment.

**Figure 52**

*Correlation Heatmap of INCIDENT_ACRES_BURNED after model optimization*



The train_test_split function from sklearn.model_selection python package was used to divide the dataset into 80/20 training-testing sub-datasets. From the xgboost python package, XGBRegressor function was used with a linear regression objective due to the quantitative nature of the outcome variable (INCIDENT_ACRES_BURNED). Similar to the prediction model for wildfire frequency, the model created for wildfire severity was optimized to obtain the most optimal parameter settings. The first model that was created with most of the parameters and the default settings produced mean-squared-error of 14.0016. After optimization, the mean-squared-error generated by the model was 12.5329 and indicates that the carbon emission drives the severity of wildfires, whereas precipitation has a negative impact on the severity of wildfire. Figure 53 shows the comparison of the severity of wildfires for the test and predicted data.

**Figure 53**

*Wildfire severity of Test and Predicted Data Comparison*

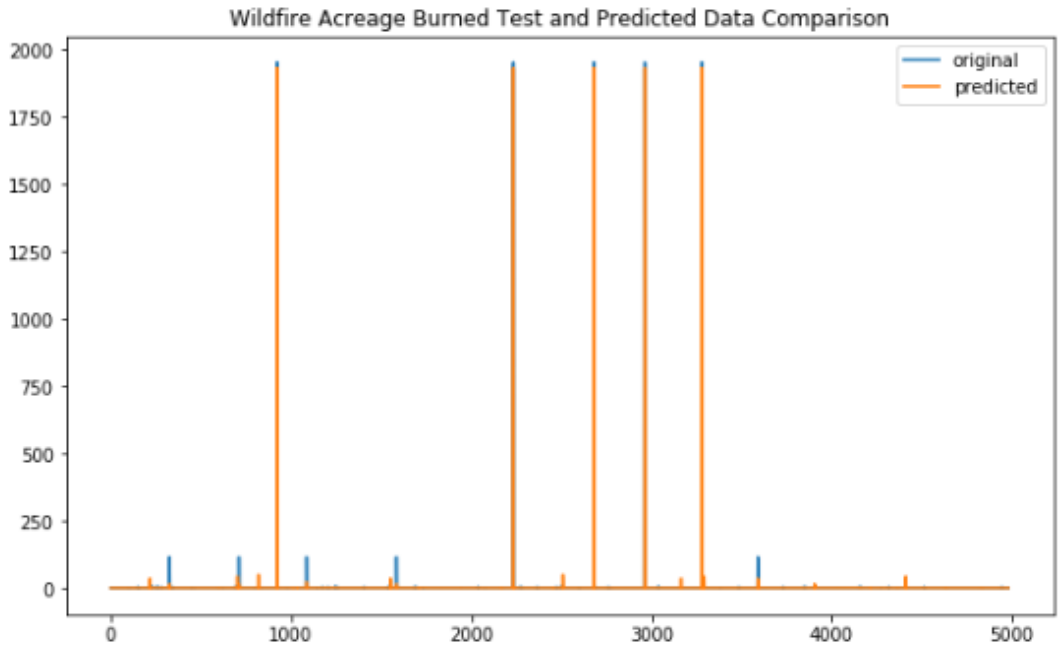


**Table 25**

*Results of XGBoost models for Wildfire frequency and severity*

**Wildfire Frequency**

| | MSE | RMSE | $R^2$ |
|---|---|---|---|
| Round 1 (default parameters) | 102.307 | 10.115 | -29.692 |
| Round 2 (gamma=0.1) | 96.866 | 9.842 | -28.06 |
| Round 3 (colsample_bytree = 0.6, gamma = 0.1) | 62.152 | 7.884 | -17.646 |
| Round 4(colsample_bytree = 0.6, gamma = 0.1, min_child_weight=1.5) | 3.852 | 1.963 | -0.155 |
| Round 5 (colsample_bytree = 0.6, gamma = 0.1, min_child_weight=1.5, n_estimators = 13) | 1.499 | 1.224 | 0.55 |
| Round 6 (colsample_bytree = 0.6, gamma = 0.1, min_child_weight=1.5, n_estimators = 13, alpha = 10) | 0.672 | 0.82 | 0.798 |

**Wildfire Severity**

| | MSE | RMSE | $R^2$ |
|---|---|---|---|
| Round 1 (all variables,, default parameters) | 14.002 | 3.742 | 0.996 |
| Round 2 (excluding CO2_EMISSIONS_TONS) | 13.151 | 3.626 | 0.997 |
| Round 3 (excluding TEMP_MAX, CO2_EMISSIONS_TONS) | 13.443 | 3.666 | 0.996 |
| Round 4 (excluding Income, TEMP_MAX, CO2_EMISSIONS_TONS) | 12.854 | 3.569 | 0.997 |
| Round 5 (excluding TEMP_MAX, (Income, TEMP_MAX, CO2_EMISSIONS_TONS) | 12.533 | 3.540 | 0.997 |

*Note:* Parameters used for training XGBoost model for wildfire severity were: colsample_bytree = 0.3, learning_rate = 0.3, gamma=0.1, max_depth = 6, alpha = 10, n_estimators = 13.

# Chapter 5 - Conclusions and Recommendations

## Conclusion

Purpose of the research was to evaluate the aspects of wildfire frequency and severity that may be influenced through carbon emissions, temperature, precipitation, and population. The results produced from the linear and logistic regression models using wildfire frequency and severity as outcome variables, and temperature, precipitation, and population as explanatory variables, concluded that the variation in both the factors of wildfires are explained by the selected explanatory variables. According to Robeson, S. M. (2015), Southern California drought has been more extreme than usual and has been thoroughly studied by the scientific community. The research supports the analysis done by Robeson S.M. (2015), that presented the effect of insufficient precipitation, extreme temperatures, and season on the frequency and severity of wildfires.

Furthermore, population has been proved to be a contributor to variation in wildfire frequency and severity. Results generated by logistic regression models support the research conducted by U.S. Fire Administration (2021), that presented the variation in frequency of wildfires beyond forested and prairie regions, as a result of growing and encroaching population of Southern California on the fringes of the wildlife areas. With the regression analyses, the research hypotheses of wildfire frequency and severity being affected by weather and populations has been accepted; however, the regression models indicated the effect of carbon emission is not significant on both frequency and severity of wildfires.

Along with multilinear regression models, the research utilized the extreme gradient boosting (XGBoost) machine learning model to answer the research question and create a prediction model for wildfire frequency and severity. The XGBoost model outperformed the

multilinear regression model producing a model with much higher prediction accuracy for both wildfire frequency and severity with the calculated mean-squared error of 0.67 and 12.53 respectively. The exceptional performance of the XGBoost model, in comparison to the multilinear regression model, lies in its machine learning algorithm, which is an ensemble learning algorithm that improves the predictability and the ability to generalize the model (Shmueli et al., 2019). The XGBoost algorithm implements a gradient tree boosting technique using a supervised learning approach (Chen & Guestrin, 2016) while focusing on the degree of error from the previous tree, leading to an improved model performance (Shmueli et al., 2019).

The process of XGBoost model optimization for wildfire frequency produced results that supported the hypothesis of this research, which stated that weather, carbon emission, and population drives the wildfire frequency and severity. The results of the XGBoost model conducted on a small data sample of 402 wildfires recorded in the month of January throughout the year 2014-2020, suggested that there is a statistically significant correlation between wildfire frequency and total fuel consumption, county, total population below poverty, maximum daily temperature, and precipitation.

The process of XGBoost model optimization for wildfire severity produced results that partially supported the hypothesis of this research. Based on the XGBoost outputs and mean-squared error, it was determined that the maximum daily temperature has no statistically significant effect on the overall model as the exclusion of this variable from the model improved the model's performance minimizing the mean-squared error from 14.00 to 13.54. Similarly, the exclusion of the carbon emission by tons improved the model's performance and decreased the mean-squared error from 13.54 to 13.44. Further exclusion of the population income and county lead to mean-squared error improvement to 12.85 and 12.53 respectively.  The precipitation and

total population showed a statistically significant effect on the finalized XGBoost model with mean-squared error of 12.53.

**Limitations**

This research included carbon emissions as one of the important factors to analyze wildfires in Southern California, however, there are other gasses emitted in the environment which may be factors for increase in temperature and wildfires in number and severity. Another data that was utilized in the research was the population data from the census bureau, and was available till the year 2020, because the latest data has not been released yet. Although the data for WUI was available, incorporating it was not feasible as part of the scope of this research. Moreover, certain aspects of weather data such as humidity, wind speed and direction, vapor pressure, solar radiation, and dew point were not included as explanatory variables to determine their effects on wildfires, as according to Alley (2018), including large number of topics in the research may have negative effects on the depth of the work.

Since the wildfire data was a secondary data, it did not include wildfires which were not reported to CalFire; the reason being that, collecting the data for non-reported wildfires requires conducting a survey within a group of people in WUI area or any residential region near forests or wildlands. According to Harris et. al. (2021), analysis of fire severity may become uncertain due to the fire suppression operation, as a result of difficulties in quantifying their effects because of rare availability of the information on when and where certain tactics were used. Given the limitations, this research is the first to analyze that there is an explicit inter-dependency cycle between wildfires and carbon emissions; however, further research is needed to accept or reject this hypothesis.

## Implications for Practitioners

The research was intended to allow practitioners (e.g., FireFighters, Paramedics, Meteorologists) the ability to predict frequency and severity of wildfires. With the earlier warning signs of wildfires, evacuations of populated areas can occur and first responders can prestage equipment in preparation of the wildfire event. This could save valuable time, money and manpower for wildfire rich environments. Another outcome of the research was the educational benefits it offered. By education the population on causes of wildfires in their area, human created wildfires can be limited or eliminated altogether.

## Recommendations and Future Scope

Adding more data sources and variables that were discussed in limitations is recommended to achieve more accurate predictions from the models, and results from analyses. Research by Harris et. al.(2021) states that additional variables such as actual evapotranspiration (AET) can be included in the research to examine wildfires, because the correlation of AET with moisture availability for plants influences fuel productivity thereby affecting the fire behavior. Similarly, the data about climatic water deficit (CWD) can be included in the research, as CWD represents drought intensity; and high-pre fire water deficits can influence fire severity because it makes trees more susceptible to mortality from fire (Harris et. al., 2021).

The research has been conducted only for the southern California region; it is recommended to replicate the analysis on the data for the Northern California region. The data collected from several sources was merged on common variables containing the name of counties and date of the observations; however, more granular data can be utilized for the research including variables such as zip codes and geographical coordinates.

## Summary

The focus of this study was to create a prediction model for the frequency and severity of wildfires in Southern California and provide the practitioners in the region with data-driven information to prepare for the potential wildfires and the accompanying destruction. The regression and XGBoost models generated concluded that weather and populations are strong predictors of wildfire frequency and severity. Although the amount of carbon emission can be used as a predictor for wildfire frequency, it may not be practical for predicting the severity of wildfires using the amount of carbon emission.

## References


2022 Fire Season Outlook. (n.d.). Retrieved from https://www.fire.ca.gov/incidents/2022/

A long view of California's climate. Study examines centuries of data to understand climate-wildfire links. (2019, May). *National Centers for Environmental Information (NCEI)*. Retrieved from https://www.ncei.noaa.gov/news/california-fire-study

Abnett, K. (2021). This Is How Much Carbon Wildfires Have Emitted This Year. World Economic Forum. Retrieved from https://www.weforum.org/agenda/2021/12/ siberia-america-wildfires-emissions-records-2021/

ArcGIS (n.d.) Wildfire Risk in the Wildland Urban Interface. *The New Normal.* Retrieved from https://www.arcgis.com/apps/Cascade/index.html?appid=cd69320c00384d8094d83b45e84fd5aa

ArcGIS Pro 3.0 (n.d. a) The architecture of a geodatabase. *esri*. Retrieved from https://pro.arcgis.com/en/pro-app/latest/help/data/geodatabases/overview/the-architecture-of-a-geodatabase.htm

ArcGIS Pro 3.0 (n.d. b) What is the Data Access module. e*sri.* Retrieved from https://pro.arcgis.com/en/pro-app/latest/arcpy/data-access/what-is-the-data-access-module-.htm

ArcMap 10.8 (2021). The geodatabase is object relational. e*sri.* Retrieved from https://desktop.arcgis.com/en/arcmap/latest/manage-data/gdb-architecture/the-geodatabase-is-object-relational.htm

Bloch, M., Reinhard, S., Tompkins, L., Pietsch, B., & Nieto del Rio, G. M. (2020, October). Fire Map: California, Oregon and Washington. The New York Times. Retrieved from https://www.nytimes.com/interactive/2020/us/fires-map-tracker.html

Brenkert-Smith, H., Champ, P.A. & Flores, N. Trying Not to Get Burned: Understanding Homeowners' Wildfire Risk–Mitigation Behaviors. *Environmental Management* 50, 1139–1151 (2012). Retrieved from https://doi.org/10.1007/s00267-012-9949-8

Brown, E. K. (March 2021). US wildfire potential: a historical view and future projection using high-resolution climate data. IOPscience. Retrieved from https://iopscience.iop.org/article/10.1088/1748-9326/aba868

CalFire. (2022). Incidents Overview. Cal Fire Department of Forestry and Fire Protection. Retrieved from https://www.fire.ca.gov/incidents/

California Air Resources Board. (2022). Frequently Asked Questions: Wildfire Emissions | California Air Resources Board. Retrieved from https://ww2.arb.ca.gov/resources/

California Drought.gov. (n.d.). Retrieved from https://www.drought.gov/states/california

Calkin, D. E., Cohen, J. D., Finney, M. A., Thompson, M. P. (2014) How risk management can prevent future wildfire disasters in the wildlandurban interface. Proc Natl Acad Sci *[PNAS]* 111(2):746–751. Retrieved from https://doi.org/10.1073/pnas.1315088111

Center for Climate Change. (2022). Wildfires and Climate Change. Retrieved from https://www.c2es.org/content/wildfires-and-climate-change/

Chen, T., & Guestrin, C. (2016, August 13). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *[C2ES]* Retrieved from https://doi.org/10.1145/2939672.2939785

CHHS Open Data Portal (2020, October 1 A). Living wage. Retrieved from https://data.chhs.ca.gov/dataset/living-wage

CHHS Open Data Portal (2020, October 1 B). Poverty Rate (<200% FPL) and Child (under 18) Poverty Rate by California Regions. Retrieved from

https://data.chhs.ca.gov/dataset/poverty-rate-by-california-regions

Chi, G., Zhu, J. Spatial Regression Models for Demographic Analysis. Popul Res Policy Rev 27, 17–42 (2008). https://doi.org/10.1007/s11113-007-9051-8

Climate Data Guide. Palmer Drought Severity Index (PDSI) | Climate Data Guide. (n.d.). Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer

Congressional Research Service (2022, November). Wildfire Statistics. InFocus. Retrieved from https://sgp.fas.org/crs/misc/IF10244.pdf

Coronel, C., & Morris, S. (2016). Database systems: design, implementation, & management. *Cengage Learning*. Available from https://books.google.com/books/about/Database_Systems_Design_Implementation_M.html?id=4JN4CgAAQBAJ

Coughlan, M. R., Ellison, A., & Cavanaugh, A. H. (2019, Fall). Social vulnerability and wildfire in the wildland-urban interface: Literature synthesis. Scholar's Bank University of Oregon. Retrieved from https://ewp.uoregon.edu/sites/ewp.uoregon.edu/files/WP_96.pdf

Dennison, P. E., Brewer, S. C., Arnold, J. D., and Moritz, M. A. (2014, April), Large wildfire trends in the western United States, 1984–2011, Geophys. Res. Lett., 41, 2928– 2933, doi:10.1002/2014GL059576/documents/frequently-asked-questions-wildfire-emissions-drought-severity-index-pdsi

Fire Safe Marin. (2021, October 6). *Fire-Hazardous Plants - Fire Safe Marin*. Fire Safe Marin - Adapt to Wildfire. Retrieved from https://firesafemarin.org/create-a-fire-smart-yard/plants/fire-hazardous-plants/

firsttuesday Journal (2021, May 2) Golden state population trends. firsttuesday Jorunal. Retrieved from https://journal.firsttuesday.us/golden-state-population-trends/9007/

Fleck, A. (2022, July 13). The Growing Danger of Californian Wildfires. Statista Infographics. Retrieved from https://www.statista.com/chart/14462/california-wildfire-deadly/

Forest Service, Bureau of Indian Affairs, Bureau of Land Management, Fish and Wildlife Service, National Park Service (2001) Urban wildland interface communities within the vicinity of federal lands that are at high risk from wildfire [Federal Register]. *National Archives.* Retrieved from https://www.federalregister.gov/documents/2001/01/04/01-52/urban-wildland-interface-communities-within-the-vicinity-offederal-lands-that-are-at-high-risk-from

FRED (2022 A). U.S. Bureau of Economic Analysis and Federal Reserve Bank of St. Louis, Per Capita Personal Income in California [CAPCPI], *Federal Reserve Bank of St. Louis*. Retrieved from https://fred.stlouisfed.org/series/CAPCPI

FRED (2022 B). U.S. Census Bureau, Resident Population in California [CAPOP], *Federal Reserve Bank of St. Louis*. Retrieved from https://fred.stlouisfed.org/series/CAPOP

Gabbe, C. J., Pierce, G., & Oxlaj, E. (2020). Subsidized Households and Wildfire Hazards in California. Environmental Management, 66(5), 873–883. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=eoah&AN=53940473&site=ehost-live&custid=natuniv

Glasmeier, A. (2022, May 10). Living Wage Calculator User's Guide/Technical Notes. Retrieved from https://livingwage.mit.edu/resources/Living-Wage-Users-Guide-Technical-Documentation-2022-05-10.pdf

Glickman, T. 2000. Glossary of Meteorology. 2d ed. Amer. Meteor. Soc., 855 pp. https://scirp.org/reference/referencespapers.aspx?referenceid=1577706

Harris, Lucas & Drury, Stacy & Farris, Calvin & Taylor, Alan. (2021, April). Prescribed fire and fire suppression operations influence wildfire severity under severe weather in Lassen Volcanic National Park, California, USA. Retrieved from : https://www.researchgate.net/publication/351154560_Prescribed_fire_and_fire_suppression_operations_influence_wildfire_severity_under_severe_weather_in_Lassen_Volcanic_National_Park_California_USA

Holden, Z. A., Swanson, A., Luce, C. H., Jolly, W. M., Maneta, M., Oyler, J. W., Warren, D. A., Parsons, R., & Affleck, D. (2018, August 20). Decreasing fire season precipitation increased recent western US forest wildfire activity. Proceedings of the National Academy of Sciences [PNAS], 115(36). Retrieved from https://doi.org/10.1073/pnas.1802316115

How to Deal with Missing Data (2022). *Master's in Data Science.* Retrieved from https://www.mastersindatascience.org/learning/how-to-deal- with-missing-data/

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021, June). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, *12*(2), 1545–1556. Retrieved from https://doi.org/10.1016/j.asej.2020.11.011

Interactive Fire and Air Quality Map. New York Times (2022). Retrieved from https://www.nytimes.com/interactive/2022/us/fire-tracker-maps.html

Jin, Y., Goulden, M. L., Faivre, N., Veraverbeke, S., Sun, F., Hall, A., Hand, M. S., Hook, S., & Randerson, J. T. (2015). Identification of two distinct fire regimes in southern california: Implications for economic impact and future change. Environmental Research Letters, 10(9). Retrieved from https://doi.org/10.1088/1748-9326/10/9/094005

Juang, C. S., Williams, A. P., Abatzoglou, J. T., Balch, J. K., Hurteau, M. D., & Moritz, M. A. (2022, March 8). Rapid Growth of Large Forest Fires Drives the Exponential Response of Annual Forest-Fire Area to Aridity in the Western United States. Geophysical Research Letters, 49(5). Retrieved from https://doi.org/10.1029/2021gl097131

Keeley, J. E., & Fotheringham, C. J. (2001). Historic fire regime in Southern California shrublands. Society for Conservation Biology, 15(6), 1536–1548. Retrieved from https://doi.org/10.1046/j.1523-1739.2001.00097.x

Keeley, J. E., & Syphard, A. D. (2021). Large California wildfires: 2020 fires in historical context. Fire Ecology, 17(1). Retrieved from https://doi.org/10.1186/s42408-021-00110-7

Keeley, J. E., Safford, H., Fotheringham, C. J., Franklin, J., & Moritz, M. (2009, September). 2007 Southern California wildfires: Lessons in complexity. OUP Academic. Retrieved from https://academic.oup.com/jof/article/107/6/287/4598876

Kerrigan, H. (2020). Settlement reached with Pg&E Wildfire victims: December 9 and 17, 2019. In H. Kerrigan (Ed.), Historic documents of 2019. CQ Press. Credo Reference: Retrieved from https://go.openathens.net/redirector/nu.edu?url=https%3A%2F%2Fsearch.credoreference.com%2Fcontent%2Fentry%2Fcqpresshd%2Fsettlement_reached_with_pgandamp_e_wildfire_victims_december_9_and_17_2019%2F0%3FinstitutionId%3D861

Kowalewski, M. (2020). Data Cleaning In 5 Easy Steps + Examples. Iterators. Retrieved from https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/

Li, S., Dao, V., Kumar, M. et al. (2022). Mapping the wildland-urban interface in California using remote sensing data. Sci Rep 12, 5789 (2022). Retrieved from https://doi.org/10.1038/s41598-022-09707-7

Liu, Z., Ciais, P., Deng, Z. et al. (2022) CarbonMonitor_method_for_website. Google Document. Retrieved from https://docs.google.com/document/d/1_q4QSUeSbwToR5ePTJnCxUzjsZFN-DbO/edit

Martinuzzi, Sebastián; Stewart, Susan I.; Helmers, David P.; Mockrin, Miranda H.; Hammer, Roger B.; Radeloff, Volker C. (2015). The 2010 wildland-urban interface of the conterminous United States - geospatial data. Fort Collins, CO: Forest Service Research Data Archive. Retrieved from https://doi.org/10.2737/RDS-2015-0012

Mensing, S. A., Michaelsen, J., & Byrne, R. (1999). A 560-year record of Santa Ana fires reconstructed from charcoal deposited in the Santa Barbara Basin, California. Quaternary Research, 51(3), 295–305. Retrieved from https://doi.org/10.1006/qres.1999.2035

Minnich, R. A. (1983). Fire mosaics in Southern California and northern Baja California. Science, 219(4590), 1287–1294. Retrieved from https://doi.org/10.1126/science.219.4590.1287

Mueller, J., Loomis, J., & González-Cabán, A. (2009). Do Repeated Wildfires Change Homebuyers' Demand for Homes in High-Risk Areas? A Hedonic Analysis of the Short and Long-Term Effects of Repeated Wildfires on House Prices in Southern California. Journal of Real Estate Finance & Economics, 38(2), 155–172. Retrieved from https://doi.org/10.1007/s11146-007-9083-1

National Oceanic and Atmospheric Administration. Homepage | *National Oceanic and Atmospheric Administration*. *US Department of Commerce.* (n.d.). Retrieved from https://www.noaa.gov/

Nauslar, N., Abatzoglou, J., & Marsh, P. (2018). The 2017 North Bay and Southern California Fires: A Case Study. Fire, 1(1), 18. MDPI AG. Retrieved from http://dx.doi.org/10.3390/fire1010018

Office of Wildland Fire (n.d.) Fuels Management. *U.S. Department of Interior*. Retrieved from https://www.doi.gov/wildlandfire/fuels#

Palmer drought severity index (PDSI). (2022). Palmer Drought Severity Index (PDSI) | NCAR - Climate Data Guide. Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi

Past weather by ZIP code - data table. NOAA Climate.gov. (n.d.). Retrieved from https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table

Pathak, T., Maskey, M., Dahlberg, J., Kearns, F., Bali, K., & Zaccaria, D. (2018). Climate change trends and impacts on California agriculture: A detailed review. Agronomy, 8(3), 25. Retrieved from https://doi.org/10.3390/agronomy8030025

Patil P. (2018, March). What is Exploratory Data Analysis? *Towards Data Science.* Retrieved from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Pausas, J.G. and Ribeiro, E. (2013). Fire and productivity. Global Ecology and Biogeography, 22: 728-736. Retrieved from https://doi.org/10.1111/geb.12043

Radeloff, V.C., Helmers, D.P., Kramer, H. A., Mockrin, M. H., Alexandre, P.M., Bar Massada, A., Butsic, V., Hawbaker, T. J., Martinuzzi, S., Syphard, A. D. (2017). The 1990–2010 wildland-urban interface of the conterminous United States - geospatial data(2nd Edition). Retrieved from https://doi.org/10.2737/RDS-2015-0012-2

Radeloff, V.C., Helmers, D.P., Kramer, H. A., Mockrin, M. H., Alexandre, P.M., Bar Massada, A., Butsic, V., Hawbaker, T. J., Martinuzzi, S., Syphard, A. D., Stewart, S. I. (2018).

Rapid growth of the US wildland-urban interface raises wildfire risk. Proc Natl Acad Sci 115(13):3314–3319. Retrieved from https://doi.org/10.1073/pnas.1718850115

Radeloff, V.C., Helmers, D.P., Mockrin, M. H., Carlson, A. R., Hawbaker, T. J., Martinuzzi, S. (2022). The 1990-2020 wildland-urban interface of the conterminous United States - geospatial data. 3rd Edition. Fort Collins, CO: Forest Service Research Data Archive. Retrieved from https://doi.org/10.2737/RDS-2015-0012-3

Rao, K., Konings, A., Yebra, M., Diffenbaugh, N., & Williams, P. (2022, Feb 07). The fastest population growth in the west's wildland fringes is in ecosystems most vulnerable to wildfires. The Conversation :Environment + Energy. Retrieved from https://www.proquest.com/newspapers/fastest-population-growth-wests-wildland-fringes/docview/2626094720/se-2

Raphael, M. N. (2003). The santa ana winds of California. *Earth Interactions, 7*(8), 1–13. Retrieved from https://doi.org/10.1175/1087-3562(2003)007<0001:tsawoc>2.0.co;2

Roberson. (2017). Wildfire Evacuation Timing in Southern California's Wildland-Urban Interface: Using Survival Analysis to Identify Differences in Resident Evacuation Behavior During the 2016 Blue Cut Fire. ProQuest Dissertations Publishing. Retrieved from https://nu.primo.exlibrisgroup.com/permalink/01NATIONAL_INST/1gmol9r/cdi_proquest_journals_1883382270

Robeson, S. M. (2015). Revisiting the recent California drought as an extreme value. Geophysical Research Letters, 42(16), 6771–6779. Retrieved from https://doi.org/10.1002/2015gl064593

Roper, W. (2020, Nov 6). States Moving Toward Clean Energy. Statista Infographics. Retrieved from https://www-statista-com.eu1.proxy.openathens .net/chart/23408/states-legislation-on-carbon-free-renewable-energy/

Ryerson, T. B., Andrews, A. E., Angevine, W. M., Bates, T. S., Brock, C. A., Cairns, B., Cohen, R. C., Cooper, O. R., de Gouw, J. A., Fehsenfeld, F. C., Ferrare, R. A., Fischer, M. L., Flagan, R. C., Goldstein, A. H., Hair, J. W., Hardesty, R. M., Hostetler, C. A., Jimenez, J. L., Langford, A. O., Wofsy, S. C. (2013). The 2010 California Research at the nexus of air quality and climate change (CalNex) field study. Journal of Geophysical Research: Atmospheres, 118(11), 5830–5866. Retrieved from https://doi.org/10.1002/jgrd.50331

Shmuel, A., & Heifetz, E. (2022, July 3). Global Wildfire Susceptibility Mapping Based on Machine Learning Models. Forests, 13(7), 1050. Retrieved from https://doi.org/10.3390/f13071050

Silvis Lab (2022). The 1990-2020 wildland-urban interface of the conterminous United States - geospatial data. University of Wisconsin-Madison. Retrieved from http://silvis.forest.wisc.edu/data/wui-change-2020/

Silvis Lab (2010). WILDLAND-URBAN INTERFACE (WUI) CHANGE 1990-2010. University of Wisconsin-Madison. Retrieved from http://silvis.forest.wisc.edu/data/wui-change/

Southern California population 2022. (n.d.). Retrieved from https://worldpopulationreview.com/regions/southern-california-population

Statista. (2022, September 2). Largest wildfires in California 2022, by acres burned. Retrieved from https://www.statista.com/statistics/943237/ leading-california-wildfires-number-acres-burned/

Statista. (2022, June 21). Number of wildfires in the U.S. 2021, by state. Retrieved from https://www.statista.com/statistics/1269724/number-of-us-wildfires-by-state/

Swain, D. L. (2021, March 9). A Shorter, Sharper Rainy Season Amplifies California Wildfire Risk. Geophysical Research Letters, 48(5). Retrieved from https://doi.org/10.1029/2021gl092843

Syphard, A. D., Brennan, T. J., & Keeley, J. E. (2017). The importance of building construction materials relative to other factors affecting structure survival during wildfire. International Journal of Disaster Risk Reduction, 21(1), 140–147. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=eoah&AN=40502360&site=ehost-live&custid=natuniv

Syphard, A. D., Keeley, J. E., Massada, A. B., Brennan, T. J., Radeloff, V. C. (2012). Housing Arrangement and Location Determine the Likelihood of Housing Loss Due to Wildfire. PLoS ONE 7(3): e33954. Retrieved from https://doi.org/10.1371/journal.pone.0033954

Telang P. (Nov 2021). What is the Objective of Exploratory Data Analysis? *Tech Canvas.* Retrieved from https://businessanalyst.techcanvass.com/objective-of- exploratory-data-analysis/

Tiseo, I. (2022, Jun 21). Global atmospheric concentration of carbon dioxide 1959-2021. Statista Infographics. Retrieved from https://www-statista-com.eu1.proxy.openathens.net/statistics/1091926/atmospheric-concentration-of-co2-historic/

U.S. Census Bureau. (2021 A). B01003 - Total Population, 1-year estimates. Retrieved from https://data.census.gov/cedsci/table?g=0400000US06_0500000US06025,06029,06037,06059,06065,06071,06073,06079,06083,06111&tid=ACSDT1Y2021.B01003

U.S. Census Bureau. (2021 B). DP05 - ACS Demographic And Housing Estimates, 1-year estimates. Retrieved from https://data.census.gov/cedsci/table?q=DP05&t=Older%20Population%3APopulations%20and%20People&g=0400000US06_0500000US06025,06029,06037,06059,06065,06071,06073,06079,06083,06111&tid=ACSDP1Y2021.DP05

U.S. Census Bureau. (2021 C). S1701 - Poverty Status In The Past 12 Months, 1-year estimates. Retrieved from https://data.census.gov/cedsci/table?g=0400000US06_0500000US06025,06029,06037,06059,06065,06071,06073,06079,06083,06111&tid=ACSST1Y2021.S1701

U.S. Census Bureau. (2022, September 15). California Remained Most Populous State but Growth Slowed Last Decade. Census.gov. Retrieved from https://www.census.gov/library/stories/state-by-state/california-population-change-between-census-decade.html

U.S. Energy Information Administration - EIA - independent statistics and analysis. Emissions by plant and by region. (2022, October). Retrieved from https://www.eia.gov/electricity/data/emissions/

U.S. Energy Information Administration - EIA - independent statistics and analysis. Energy-Related Carbon Dioxide Emissions by State, 2005 - 2016. (2019, February). Retrieved from https://www.eia.gov/environment/emissions/state/analysis/

U.S. Fire Administration. (2021). Analysis of NFIRS Incidents in the Wildland Urban Interface An Analysis of California NFIRS Data, 2009-2011. Retrieved from https://www.usfa.fema.gov/wui/data/wildfire-report-series/nfirs-wui-incidents-california.html

U.S. Fire Administrator. (2022). An Analysis of NFIRS Data for Selected Wildfires Including Impacts in Wildland Urban Interface Areas. Retrieved from https://www.usfa.fema.gov/wui/data/wildfire-report-series/nfirs-data-for-selected-wildfires.html

US Carbon Monitor. Carbon monitor. (n.d.). Retrieved from https://us.carbonmonitor.org/

Westerling, A. L., Cayan, D. R., Brown, T. J., Hall, B. L., & Riddle, L. G. (2004). Climate, Santa Ana winds and autumn wildfires in Southern California. Eos, Transactions American Geophysical Union, 85(31), 289. Retrieved from https://doi.org/10.1029/2004eo310001

Western Fire Chief Association. (2022), California Fire Season: In-Depth Guide. Retrieved from https://wfca.com/articles/california-fire-season-in-depth-guide/

WFIGS - Wildland Fire Locations Full History. (n.d.). Retrieved from https://data-nifc.opendata.arcgis.com/datasets/nifc::wfigs-wildland-fire-locations-full-history/explore?location=-0.000000,0.000000,2.37

Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch, J. K., & Lettenmaier, D. P. (2019, August). Observed Impacts of Anthropogenic Climate Change on Wildfire in California. Earth's Future, 7(8), 892–910. Retrieved from https://doi.org/10.1029/2019ef001210