**TEMASEK POLYTECHNIC   
SCHOOL OF ENGINEERING**

**Diploma in [Computer Engineering]**

**[DVISANLY Project Report]**

**Submitted By**

**[Lim Kim Ze Michell]**

**Class: PE06**

**Declaration of Originality**

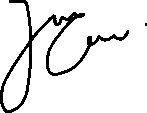
I am the originator of this work and I have appropriately acknowledged all other original sources used as my reference for this work.

I understand that Plagiarism is the act of taking and using the whole or any part of another person’s work, including work generated by AI, and presenting it as my own.

I understand that Plagiarism is an academic offence

and if I am found to have committed or abetted the offence of plagiarism in relation to this submitted work, disciplinary action will be enforced.

**Student’s Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



**AY2023/2024 APRSEMESTER**

**Declaration on the use of Generative AI tools for assignments**

*This section is* ***only applicable if the subject explicitly allows students to use Generative AI tools.*** *Students are to provide evidence on how they have used the tools to help draft their assignments.*

|  |
| --- |
| Describe how you have used Generative AI tools such as ChatGPT or Dall.E-2 in your assignment.  Show snapshots of the conversations with the AI tool (i.e., the prompts you used and the response you get from the AI tool). |
|  |
| How to indicate the reference?  The content generated by AI tools are not retrievable except by the user who generated them, so they are considered non-recoverable sources. For non-recoverable sources:   * do not include in a reference list * cite within the text as personal communications or correspondence   Based on APA 7th edition referencing format,  (Communicator, personal communication, Month Day, year)  E.g. (Paraphrase from OpenAI's ChatGPT AI language model, personal communication, March 8­, 2023). |

**Important Note:**

* Do not copy answers produced by the AI tool in totality as it is considered as plagiarism.
* Do not rely on any information produced by the AI tool blindly. You should always verify the answer with other sources. Do not assume that these answers provided by the AI tool are correct.
* To achieve quality outputs from the AI tool, you should provide good prompt that is clear and specific. Be precise and provide context. Avoid asking open-ended questions.

Content Page

1. Pre-Project Plan
2. Preliminary Questions
3. Monitor
4. Introduction
5. Data Cleaning
6. Exploratory Data Analysis
7. Answer to Preliminary questions

1. Further Insights Questions and answer
2. Data Modeling
3. Conclusion
4. Reflection
5. References (if any)

## Pre-Project Plan

|  |
| --- |
| **Goal Setting** |
| I aim to complete my project by *16/06/2023.*  I shall take initiative to find out the information needed.  I shall check the project rubric to ensure all items are done before submission. |

**My data set is** \_\_\_Apr23 WildFires\_Data\_1\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Preliminary questions**

**My preliminary questions that I will answer from my data set:**

**1. Is there a seasonal pattern in occurrences of wildfire throughout the year?**

**2. a. Which states have the highest percentage of man-made compared to natural fires?**

**b. In the top 3 states, which manmade causes had the highest percentage of occurrence?**

**3. How does the average fire size change over the years?**

**4. Is there a relationship between cause of fire and fire size?**

**5. Are there any trends in the cause of wildfires over the years?**

**6. a. Are there any hotspots of wildfires based on geographical coordinates?**

**b. Where are the top 3 countries with highest wildfire counts located on the map?**

**7. a. How does frequency of wildfires differ between each reporting agency?**

**b. How do they differ between each reporting agency Unit IDs?**

1. **Monitor**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task/Milestone** | **By When** | **Actual Completed Date** | **Comment**  (on-time/delay/early) |
| Download the data.  Understand the rows and columns. | **1 Jun 2023** | **1 Jun 2023** | **On-time** |
| Background research of delivery mode, function of eCommerce. | **6 Jun 2023** | **2 Jun 2023** | **Early** |
| Perform data cleaning. | **10 Jun 2023** | **10 Jun 2023** | **On-time** |
| Perform data transformation. | **11 Jun 2023** | **11 Jun 2023** | **On-time** |
| Exploratory Data Analysis | **16 Jun 2023** | **16 Jun 2023** | **On-time** |
| Submit Report 1 | **16Jun 2023**  **(Due date)** | **16 Jun 2023** | **On-time** |
| Answer my preliminary questions | **28 Jul 2023** | **28 Jul 2023** | **On-time** |
| Data modelling | **28 Jul 2023** | **28 Jul 2023** | **On-time** |
| Final report conclusion and reflection | **29 Jul 2023** | **29 Jul 2023** | **On-time** |
| Create Dashboard | **30 Jul 2023** | **30 Jul 2023** | **On-time** |

1. **Introduction**

A wildfire is an uncontrolled fire that burns in wildland vegetation, often in rural areas. Wildfires can burn in forests, grasslands, savannas, and other ecosystems, and have been doing so for hundreds of millions of years. They are not limited to a particular continent or environment. [1]

Wildfires can start from natural causes, such as lightning, but they are usually caused by humans, such as campers or hikers who did not put out their campfire properly. They spread quickly and can also affect natural sources such as soil, animals, forests, destroy homes, and put people’s lives in danger. [2]

Wildfires are increasing around the globe and the risk grows in extremely dry conditions such as drought, heat waves and during high winds. Wildfire events are getting more extreme with climate change leading to warmer temperatures and drier conditions. [3]

Wildfire data is collected by sensors aboard orbiting satellites, carried aboard aircraft, or installed on the ground which can be used to assess conditions before a burn, track fire size in near real-time, and assess the environmental impact of an historic burn. [4] These data are then used to generate consistent national data for risk assessment, planning, budget formulation, predicting trends and decision support at multiple scales. [5]

Source System Type (SOURCE\_SYSTEM\_TYPE) refers to the type of source database or system the record was drawn from. These systems provide access from the SAP BW (Business Information Warehouse). [6]

NWCG Reporting Agency (NWCG\_REPORTING\_AGENCY) refers to Active National Wildfire Coordinating Group. The group provides national leadership to enable interoperable wildland fire operations among federal, state, local, Tribal, and territorial partners. One of their primary objectives is to ensure is to ensure all NWCG activities contribute to safe, effective, and coordinated national interagency wildland fire operations. [7]

NWCG Reporting Unit ID (NWCG\_REPORTING\_UNIT\_ID) are fundamental elements for the interagency incident management community because they provide a common interagency identification of participating organizations and the roles they play. Unit IDs and its associated information are the standard way the organizational units communicate who they are, report incidents they are hosting and account for resources they provide. [8]

Fire year (FIRE\_YEAR) refers to the calendar year in which the fire was discovered or confirmed to exist. This indicator tracks the frequency of all reported wildfires every year which can be used to show trend in occurrences of wildfires during the period.

Statistical cause description (STAT\_CAUSE\_DESCR) refers to the cause of the wildfire. Understanding the cause of a wildfire can help to identify factors that contribute to its ignition. Through this, preventive measures can be taken to reduce the risk of similar fires in the future.

Fire size class (FIRE\_SIZE\_CLASS) is a code that corresponds to one of the several ranges of fire size based on the number of acres within the fire perimeter. [9] Classification provides a simpler and more convenient overview of how frequently a fire size class occurs.

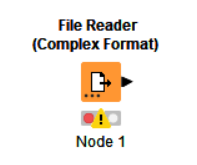
Owner description OWNER\_DESCR) refers to the name of the primary owner or entity responsible for managing the land at the point of fire origin. Different landowners have varying management goals and knowing about different land ownerships is relevant to inform perspectives on land management among public and private entities. [10]

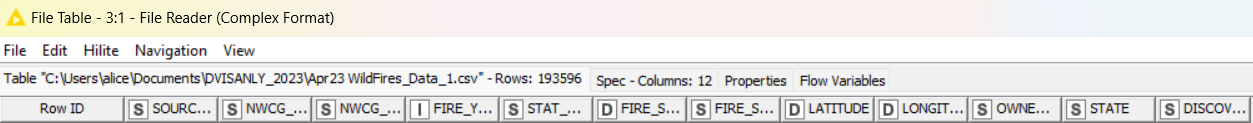
State (STATE) in the USA where fire occurs at the most and least can help in resource allocation, enabling a more efficient deployment of personnel to the affected area. California is the state with the highest risk of wildfires, with the top 5 costliest wildfires occurrences, each causing several billions of dollars in insured losses. [11] By knowing this, awareness can be raised to the public to inform and educate them on the risks and dangers of wildfire. Likewise, more resources can be allocated to this state, and research data from past burns can be collected and used to predict costs of future burns.

Discovery date (DISCOVERY\_DATE\_2) refers to the date on which the fire was discovered or confirmed to exist. Some fires are seasonal fires, while others occur in a fire regime. Identifying the months and years in which the fire occurs can aid in the prediction of future burns at the specific area.

Latitude (LATITUDE) and longitude (LONGITUDE) are a simultaneous pair of coordinates to help locate the point location of the fire. Topography affects wildfire, which spread quickly downhill and slowly uphill. Varying fuels surrounding the different areas such as dried grass, organic matter and growing vegetation are vital as they have different speeds of ignition and extinguish rate. More often, wildfires occur in dry, hot and windy weather conditions, which are common for countries located near the equator. Sometimes, wildfires are intentionally set to decrease fuels in a given area to promote the health of fire-adapted ecosystems. [12]

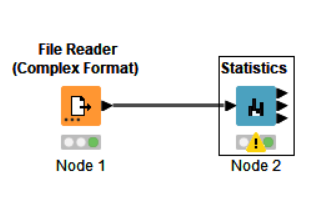
In order to learn more about the variables in the data such as missing values and variable types, we have to use File Reader (Complex Format) (Node 1) to read the given data from the csv data file.





**There are 193596 observations and 11 variables (not including discovery date).**

After configuring and selecting the file to be used, Statistics (Node 2) is used to generate numerical summaries for the variable in the given dataset.



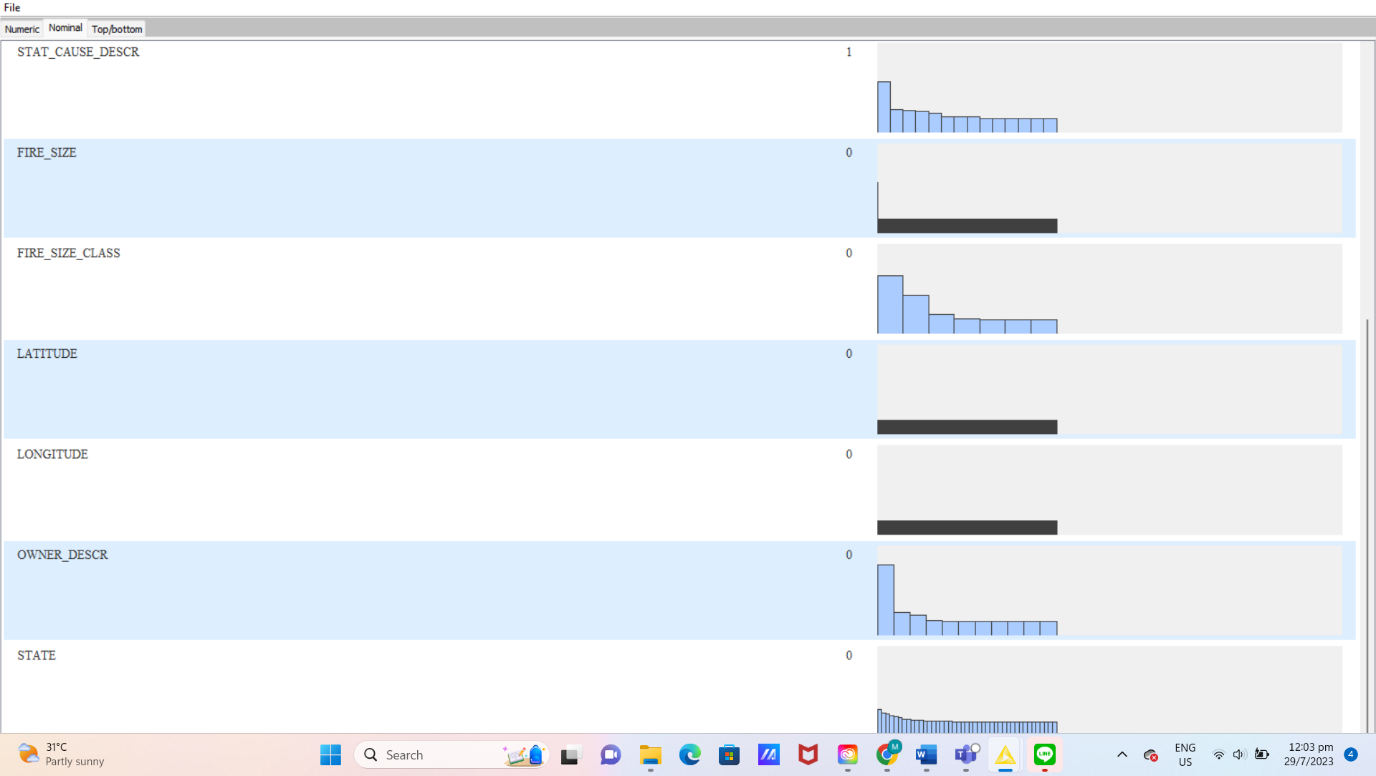
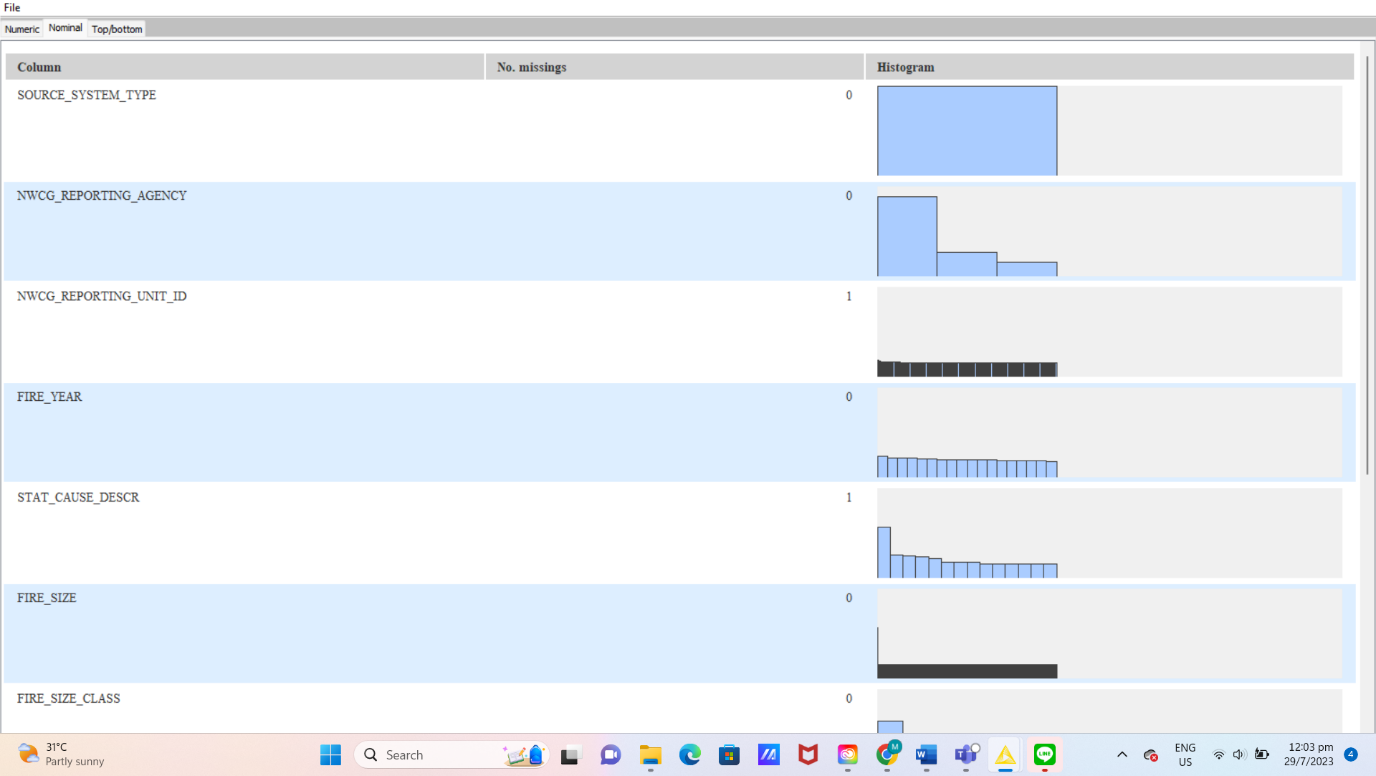
|  |  |
| --- | --- |
| **Name of variable** | **Data type** |
| SOURCE\_SYSTEM\_TYPE | Categorical: String: FED |
| NWCG\_REPORTING\_AGENCY | Categorical: String: FS, BIA, TRIBE |
| NWCG\_REPORTING\_UNIT\_ID | Categorical: String: USAZCOF, USIDNCF, USAZTNF, USAZASF, USNMGNF, USORDEF, USCAANF, USAZKNF, USNVHTF, USWAOWF, etc. (There are 188 Unit IDs) |
| FIRE\_YEAR | Numerical: Discrete: 1992 to 2009 |
| STAT\_CAUSE\_DESCR | Categorical: String: Arson, Campfire, Children, Debris Burning, Equipment Use, Fireworks, Lightning, Miscellaneous, Missing/Undefined, Powerline, Railroad, Smoking, Structure |
| FIRE\_SIZE | Numerical: Continuous: 0.01 to 499945 |
| FIRE\_SIZE\_CLASS | Categorical: String: A, B, C, D, E, F, G |
| LATITUDE | Numerical: Continuous: 26.1 to 66.833 |
| LONGITUDE | Numerical: Continuous: -163.786 to -67.066 |
| OWNER\_DESCR | Categorical: String: USFS, BIA, STATE OR PRIVATE, MISSING/NOT SPECIFIED, OTHER FEDERAL, PRIVATE, STATE, TRIBAL, BLM, FWS, NPS |
| STATE | Categorical: String: AK, AL, AR, AZ, CA, CO, FL, GA, IA, ID, IL, IN, KS, KY, LA, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NM, NV, NY, OH, OK, OR, PA, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY |

1. **Data Cleaning**

From the Statistics Table, we can see that there are missing values in the variables.

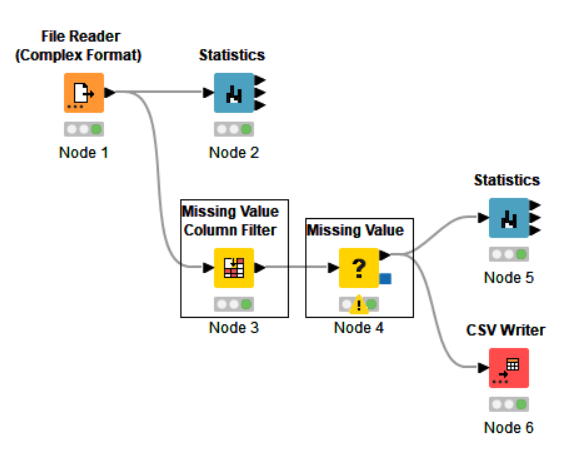
There are:

* 1 missing value for NWCG\_REPORTING\_UNIT\_ID
* 1 missing value for STAT\_CAUSE\_DESCR



To remove all the missing values in the data tables so that there will not be disruption or inconsistencies in our analysis, we can remove the rows containing missing values.

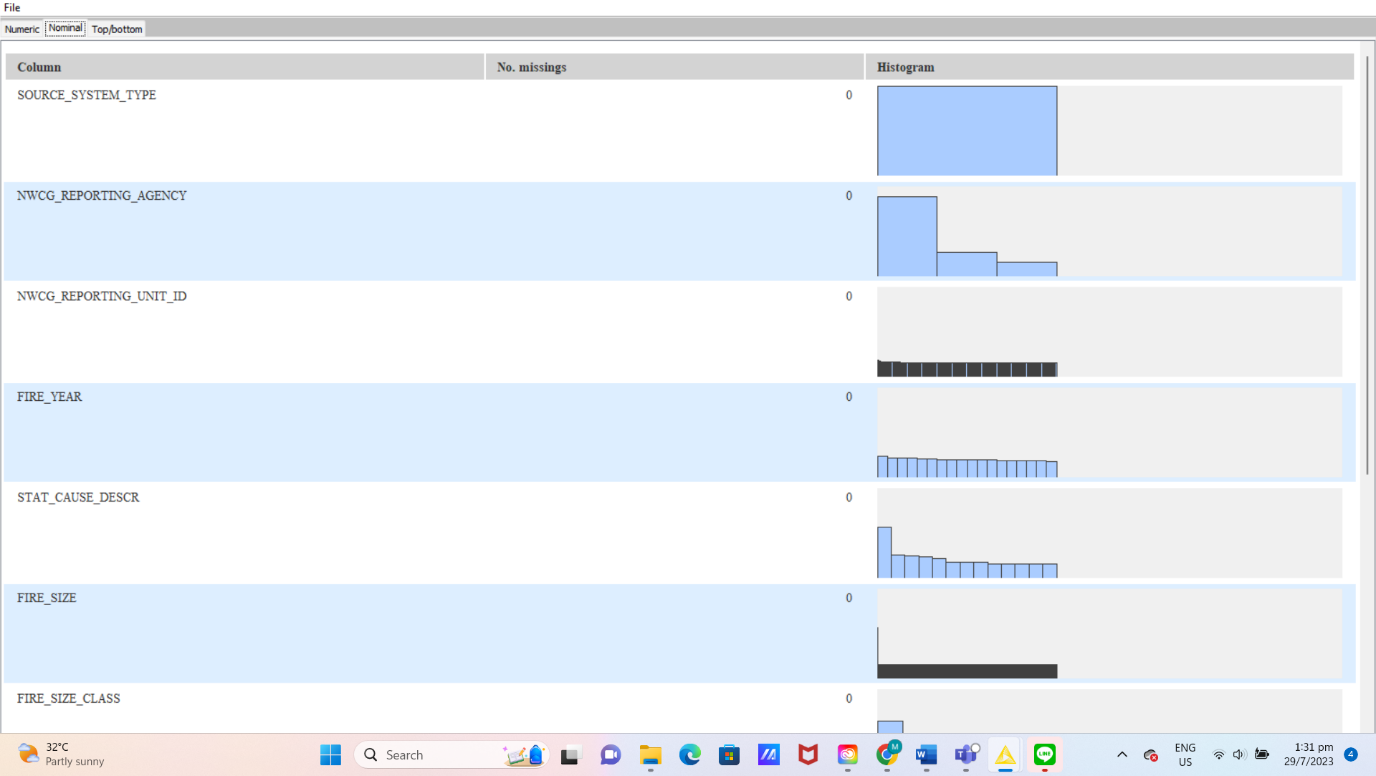
1. Using the Missing Value Column Filter (Node 3), if the missing value of a variable is equal to or exceeds 90%, the column of the variable will be removed.
2. Using the Missing Value (Node 4), if there is a missing value in the column, the row of the occurrence will be removed, ensuring the dataset will not have any missing values to disrupt the analysis.

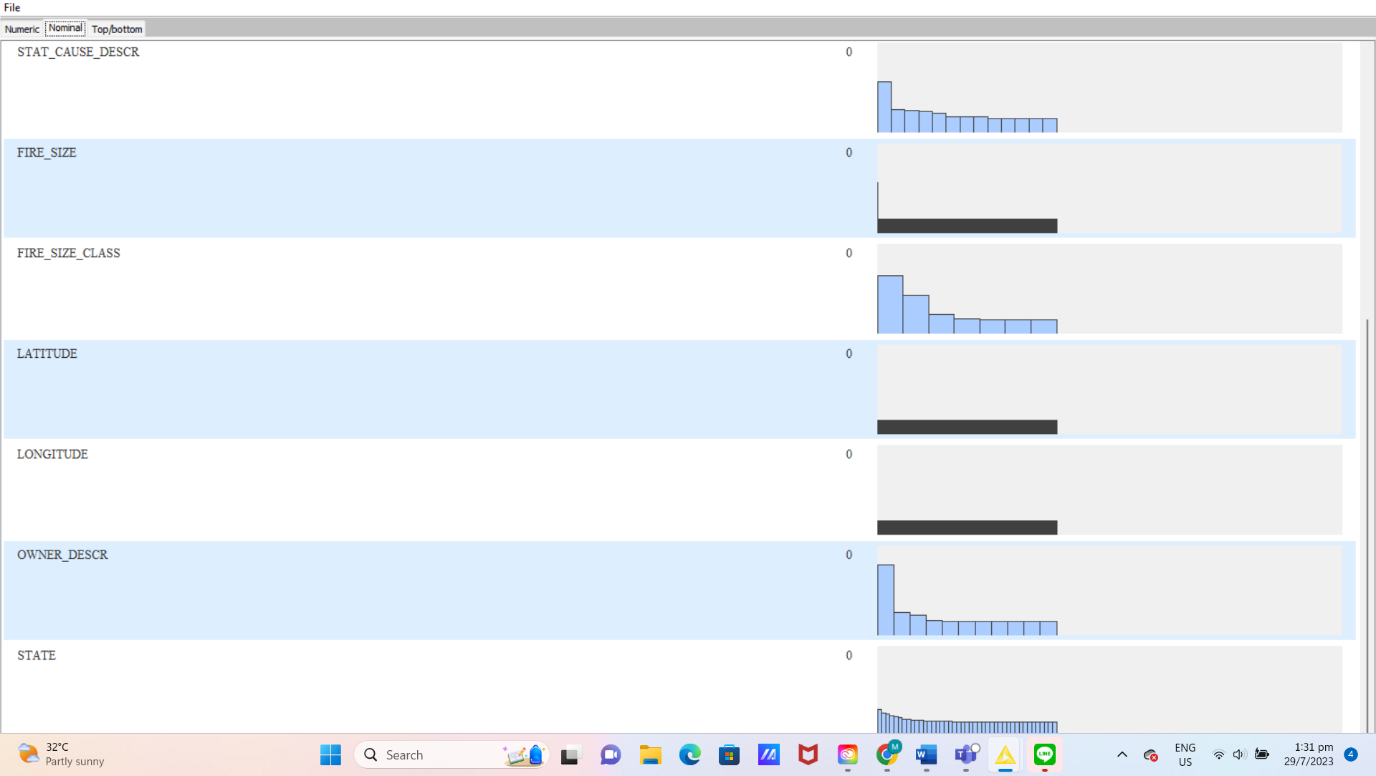


**After cleaning the dataset, there are 193594 observations and 11 variables (not including discovery date).**

****

Using the Statistics (Node 5) to view the cleaned dataset, there are now 0 missing values in every column. This means that the dataset is cleaned and completed for more accurate exploration and analysis.





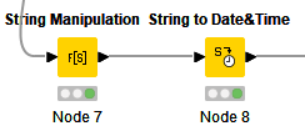
Using the CSV Writer (Node 6), the cleaned data set can be saved for further exploration. It is named as Apr23 WildFires\_Data\_1\_Cleaned.

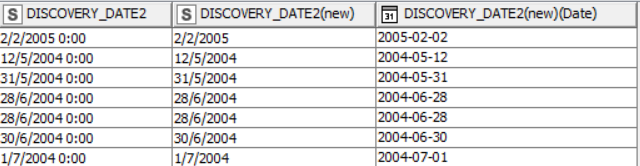
If there is wrong data type present in the data, nodes such as “String to Date&Time” can be used to convert string data to Date & Time format.

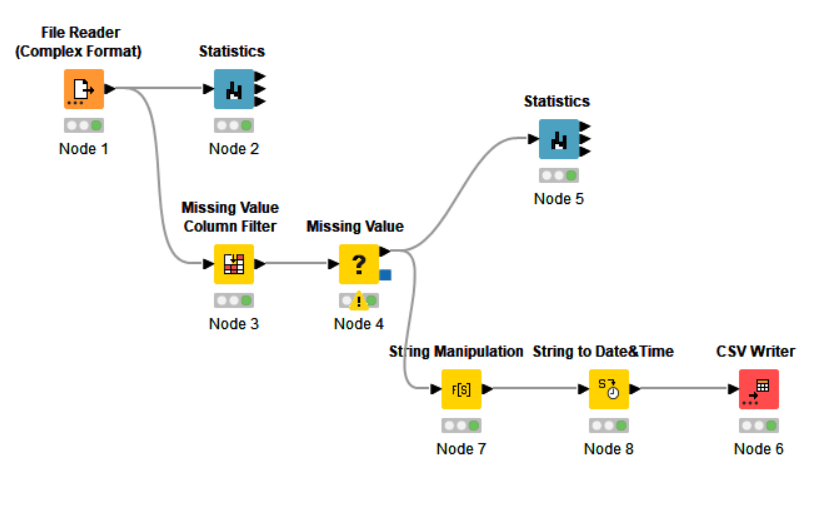
If there are unnecessary characters in variable, nodes such as “String Manipulation” can be used to remove or replace specified characters in the string.

Using String Manipulation (Node 7), time data from the DISCOVERY\_DATE\_2 variable is removed because all time is the same at 0:00. Since time may not be applicable as there are no variations, this data may not be useful or significant. The new column consisting of only string type date data for DISCOVERY\_DATE\_2 is replaced.

Using String to Date&Time (Node 8), string type data from DISCOVERY\_DATE2 variable is converted into a suitable date format. This new set of data replaces the column for the variable DISCOVERY\_DATE2 in date type format.

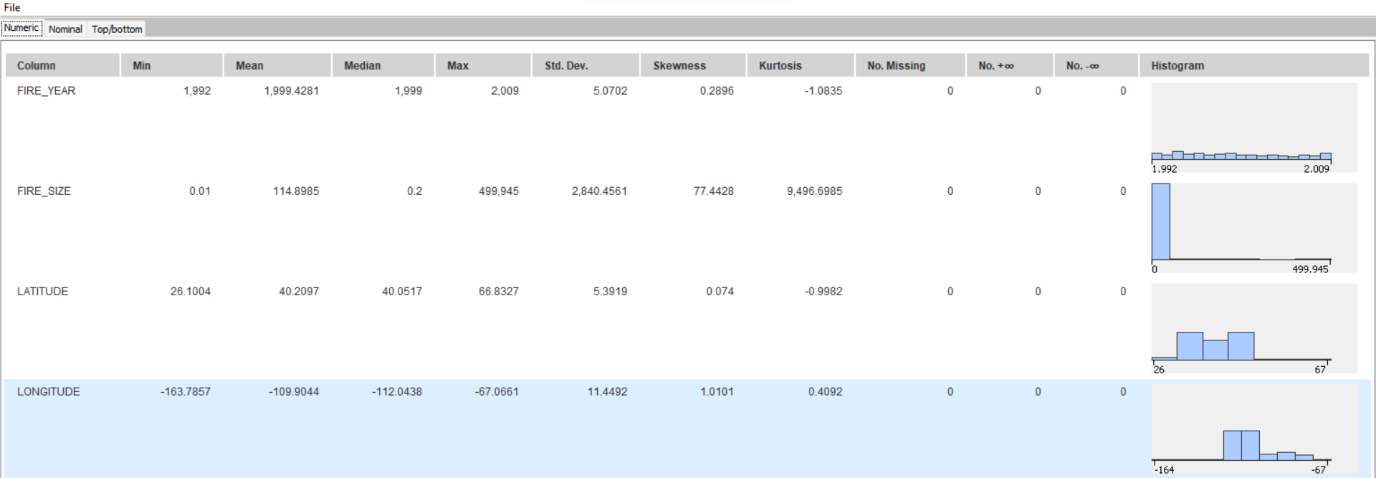
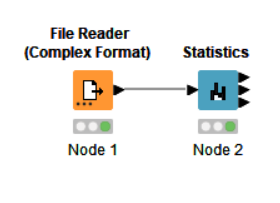




Final workflow: 

1. **Exploratory Data Analysis**

**6.1 Statistics Results**



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Min | Max | Mean | Standard Deviation | Median |
| FIRE\_YEAR | 1992 | 2009 | 1999.4281 | 5.0702 | 1999 |
| FIRE\_SIZE | 0.01 | 499945 | 114.8985 | 2840.4561 | 0.2 |
| LATITUDE | 26.1004 | 66.8327 | 40.2097 | 5.3919 | 40.0517 |
| LONGITUDE | -163.7857 | -67.0661 | -109.9044 | 11.4492 | -112.0438 |

**6.2 Correlations (Linear Correlation Matrix and Scatter Plot)**

A diagram of a computer program

Description automatically generated

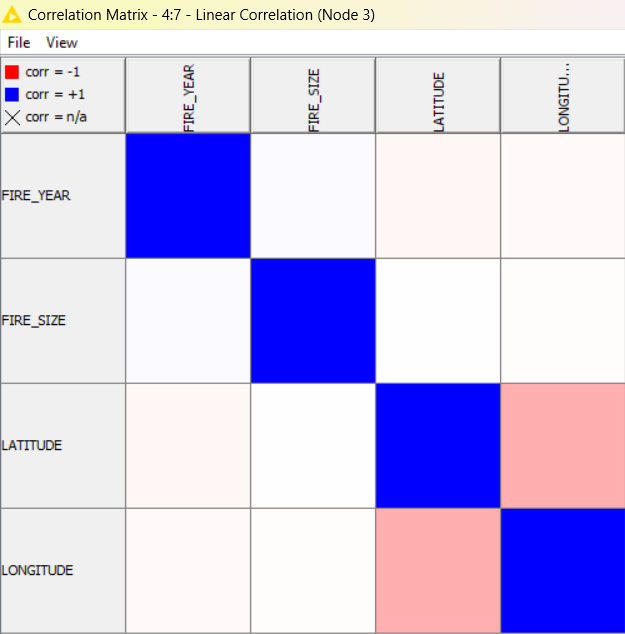


Figure 1: Linear Correlations with Knime

Using the Linear Correlation (Node 3), the strength of correlation between each variable is shown in a correlation matrix in Figure 1. (Note that longitude and latitude are a pair of coordinates on the map, thus will not be included in correlation analysis against each other.)

Figure 1 shows that:

* Both longitude and latitude have a very weak negative correlation with fire year with R-value of -0.0244 and -0.0358 respectively.
* Both longitude and latitude also have very weak correlations with fire size. Longitude has a negative R-value of 0.0118 and latitude has a positive R-value of 0.0021.



Figure 2: Linear Correlations with Tableau

Similarly, Tableau also shows the same correlations on a scatter plot,

* Both longitude and latitude have a very weak correlation with fire year.
* Longitude has a very weak negative correlation with fire size.
* Latitude has a very weak positive correlation with fire size.

**6.3 Mode of Reporting Agency (Bar Chart)**

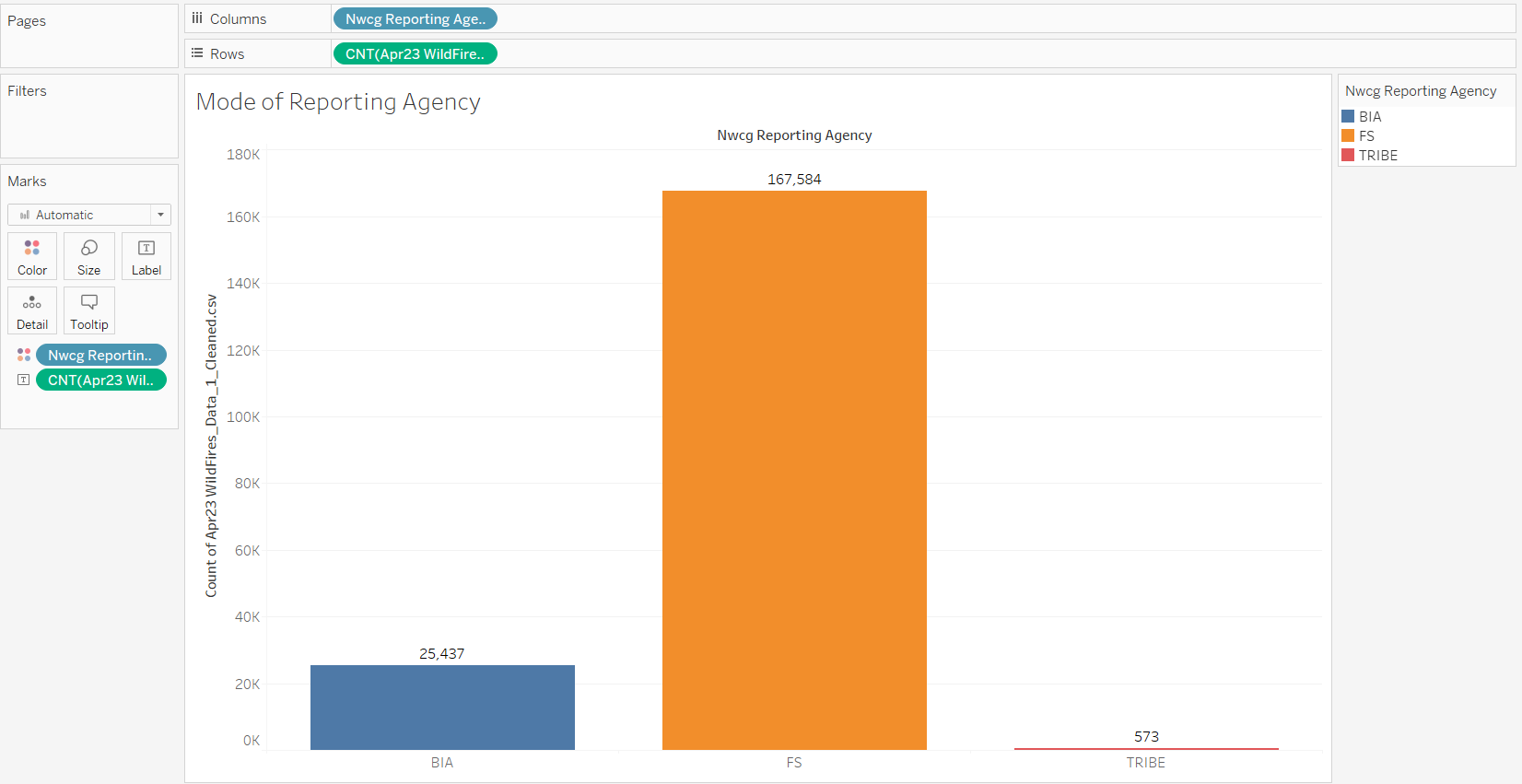
****

Figure 3: Mode of NWCG Reporting Agency

The mode of NWCG Reporting Agency is FS (Forest Service) with the highest count of 167,584 wildfire occurrences. This suggests forests are more prone to fires as the Forest Service protects and manages forested areas.

**6.****4 Percentage of Fire Occurrences by NWCG Reporting Agency (Pie Chart)**

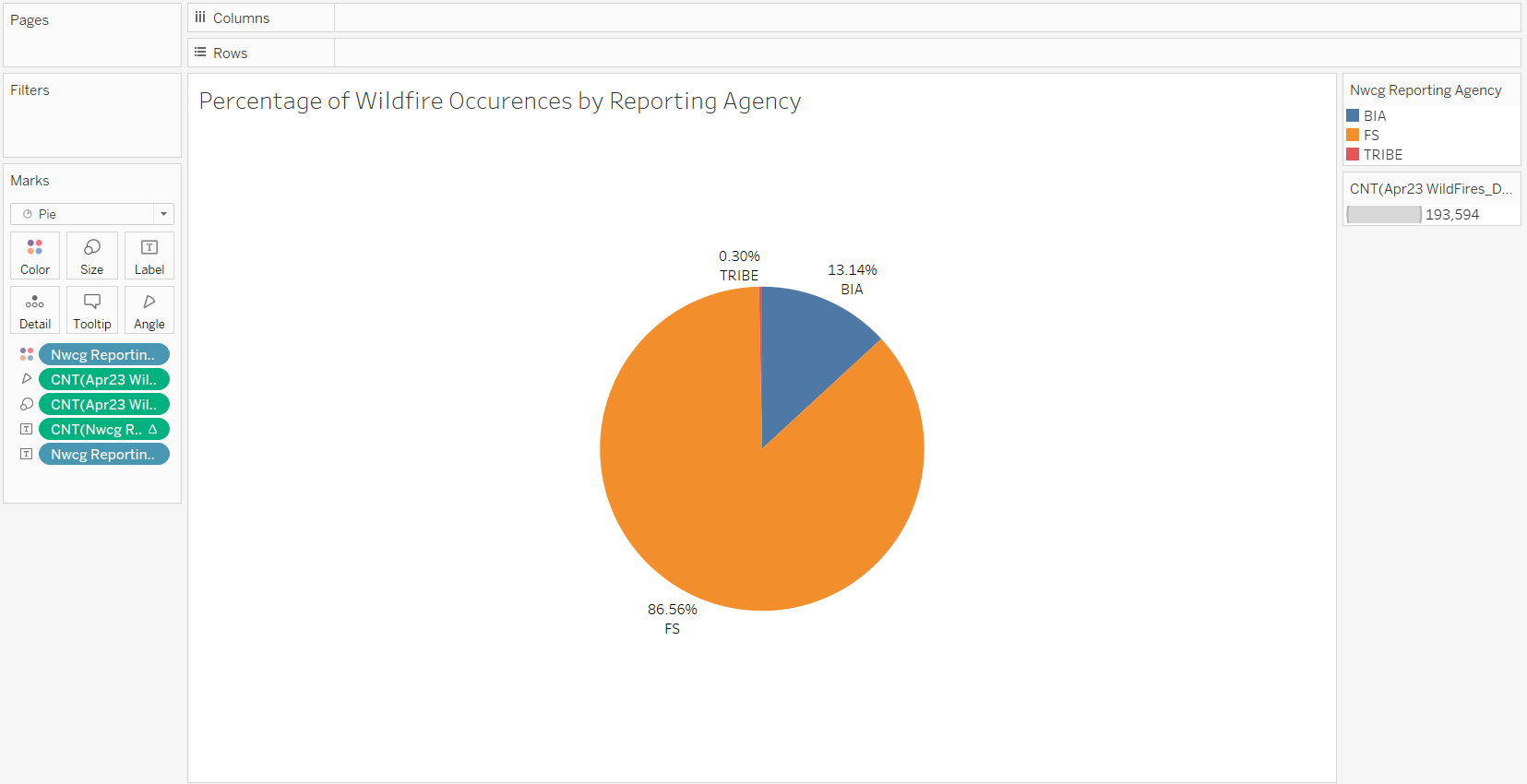


Figure 4: Percentage of Fire Occurrences by NWCG Reporting Agency

Similarly, the pie chart in Figure 4 also shows that FS (Forest Service) has the highest percentage of wildfire occurrences of 86.56%. BIA (Bureau of Indian Affairs) comes in next with 13.14%, followed by Tribe with 0.30%. This also suggests that fires occur least in tribes, hence tribes are successful in managing wildfires.

**6.5 Mode of NWCG Reporting Unit ID (Bar Chart)**

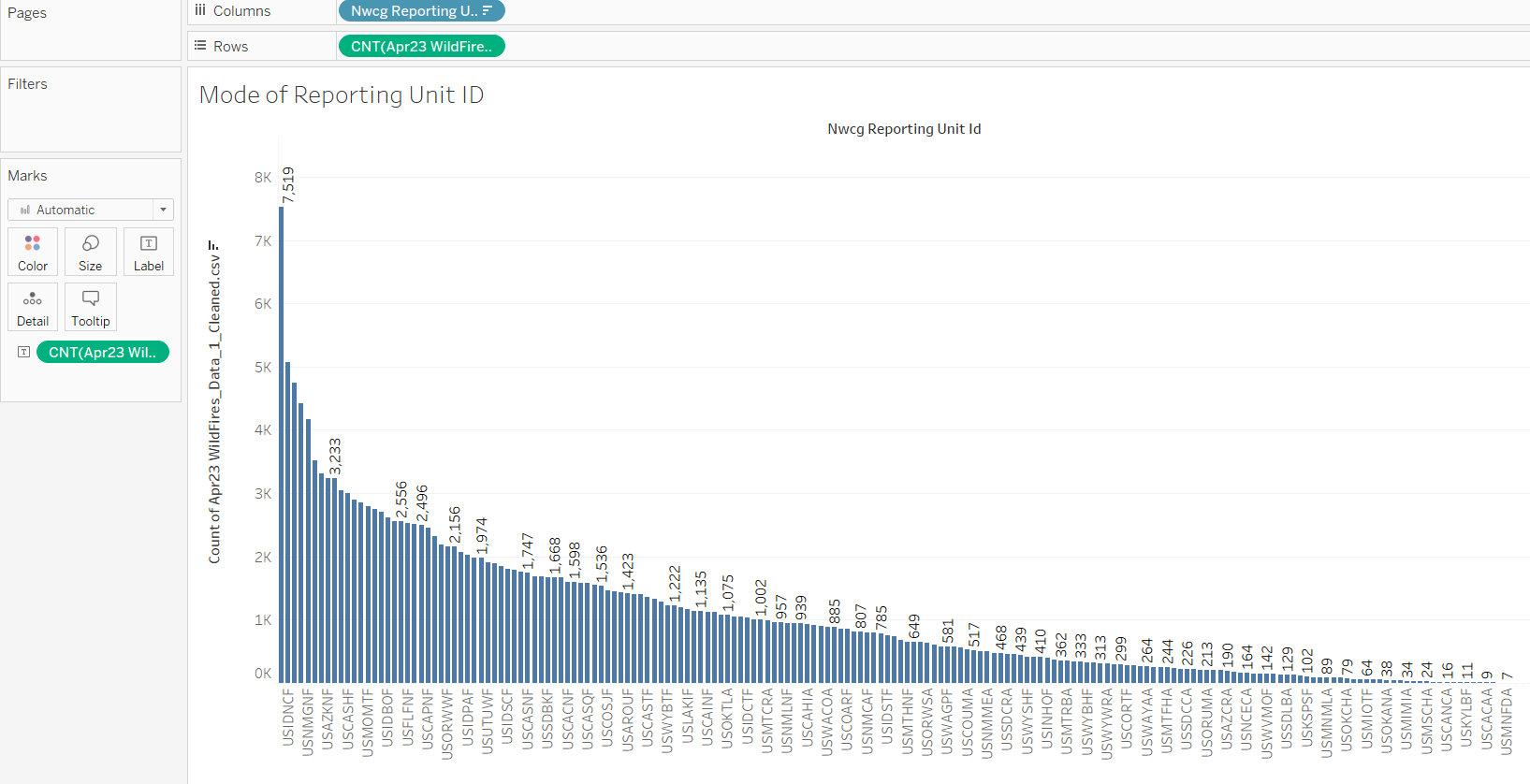


Figure 5: Mode of NWCG Reporting Unit ID

The mode of NWCG Reporting Unit ID is USAZCOF with a count of 7519. This could suggest more resources need to be provided to this unit as there are many wildfires reported by USAZCOF.

**6.6 Statistics of Fire Size (Box Plot)**

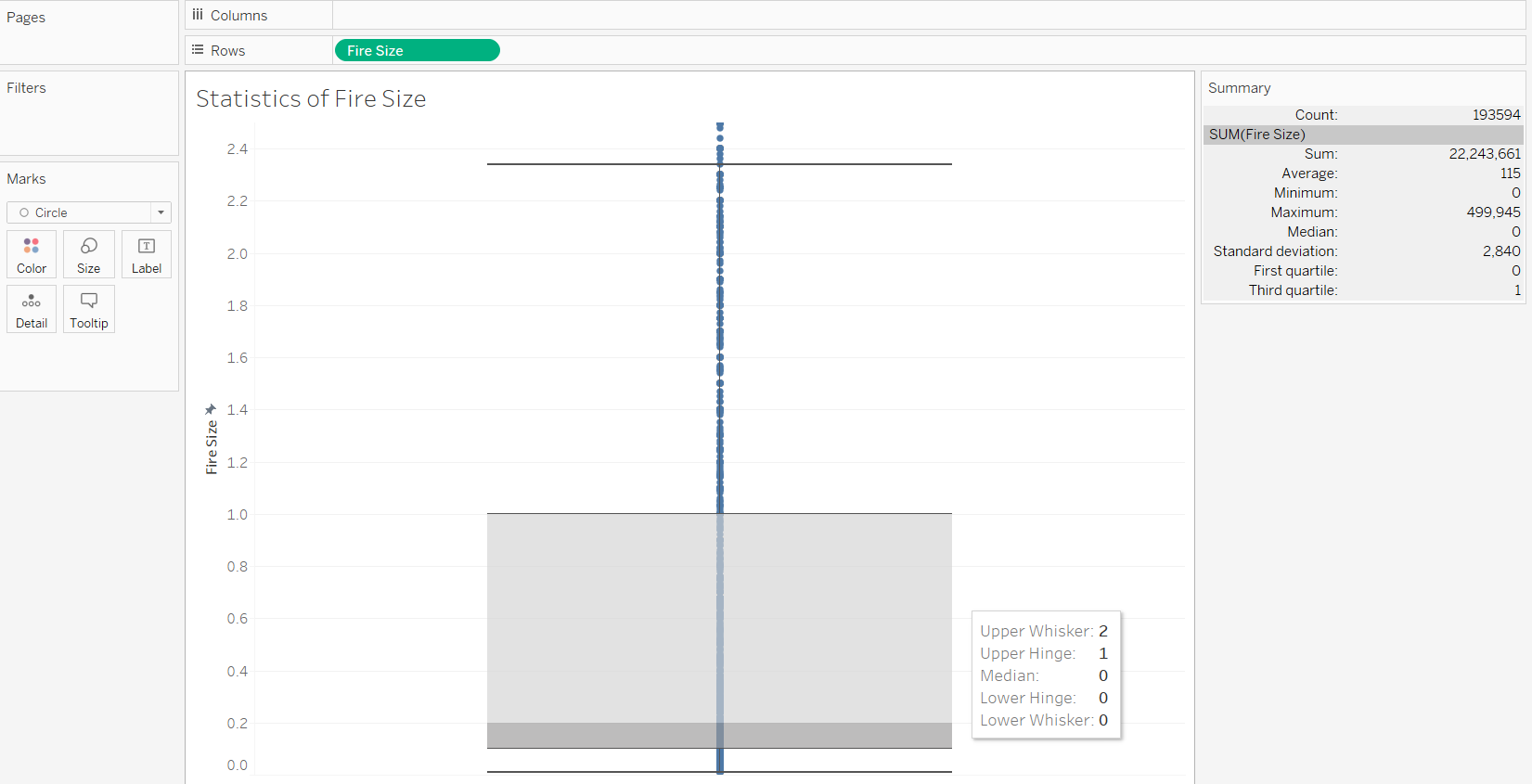


Figure 6: Close Up on Statistics of Fire Size

The median for fire size is 0.2, which is very small. This means that 50% of all fire occurrences are less than 0.2 acres wide. The lower whisker for fire size seems to be 0 as the value has been truncated to 1 decimal place. The upper whisker for fire size is around 2.4. The interquartile range is 1 which is also very small.



Figure 7: Overall Statistics of Fire Size

The maximum value for fire size is enormous at 499,945. Since the bin sizes have become significantly larger once exposed to full view, and maximum value for fire size is significantly higher than the upper whisker value, this suggests that the box plot is affected by many outliers.

**6.7 Mode of Fire Size (Histogram)**

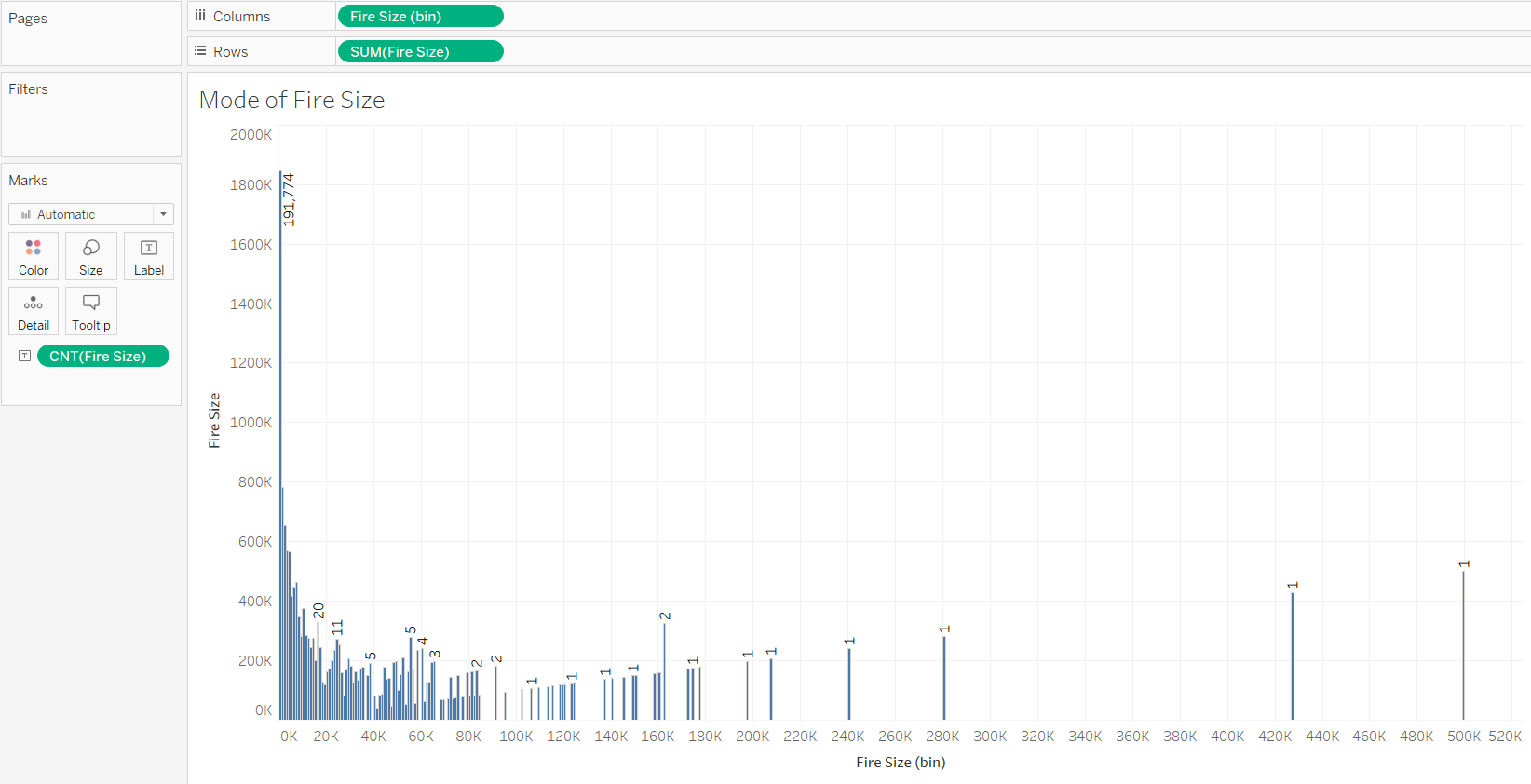
****

Figure 8: Overall Mode of Fire Size

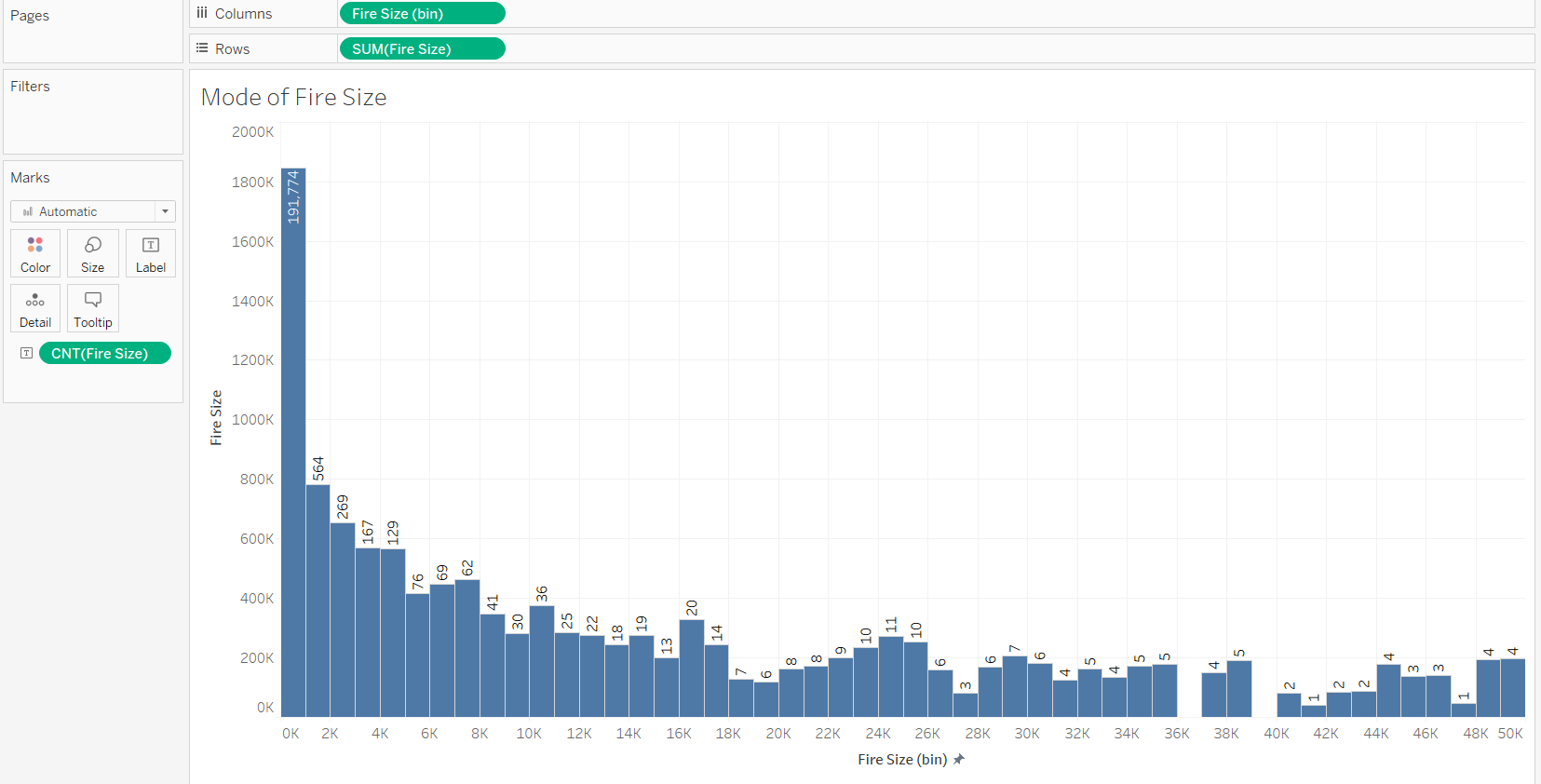


Figure 9: Close Up on Mode of Fire Size

There is an extremely wide spread of data ranging from 0k to 500k. Closing up on the values, the mode of fire sizes occurs between 0k to 1k, with the highest count at 193228. This suggests most wildfires are containable as they occur within have small fire sizes occurring within a small area.

**6.8 Mode of Fire Year (Bar Chart)**

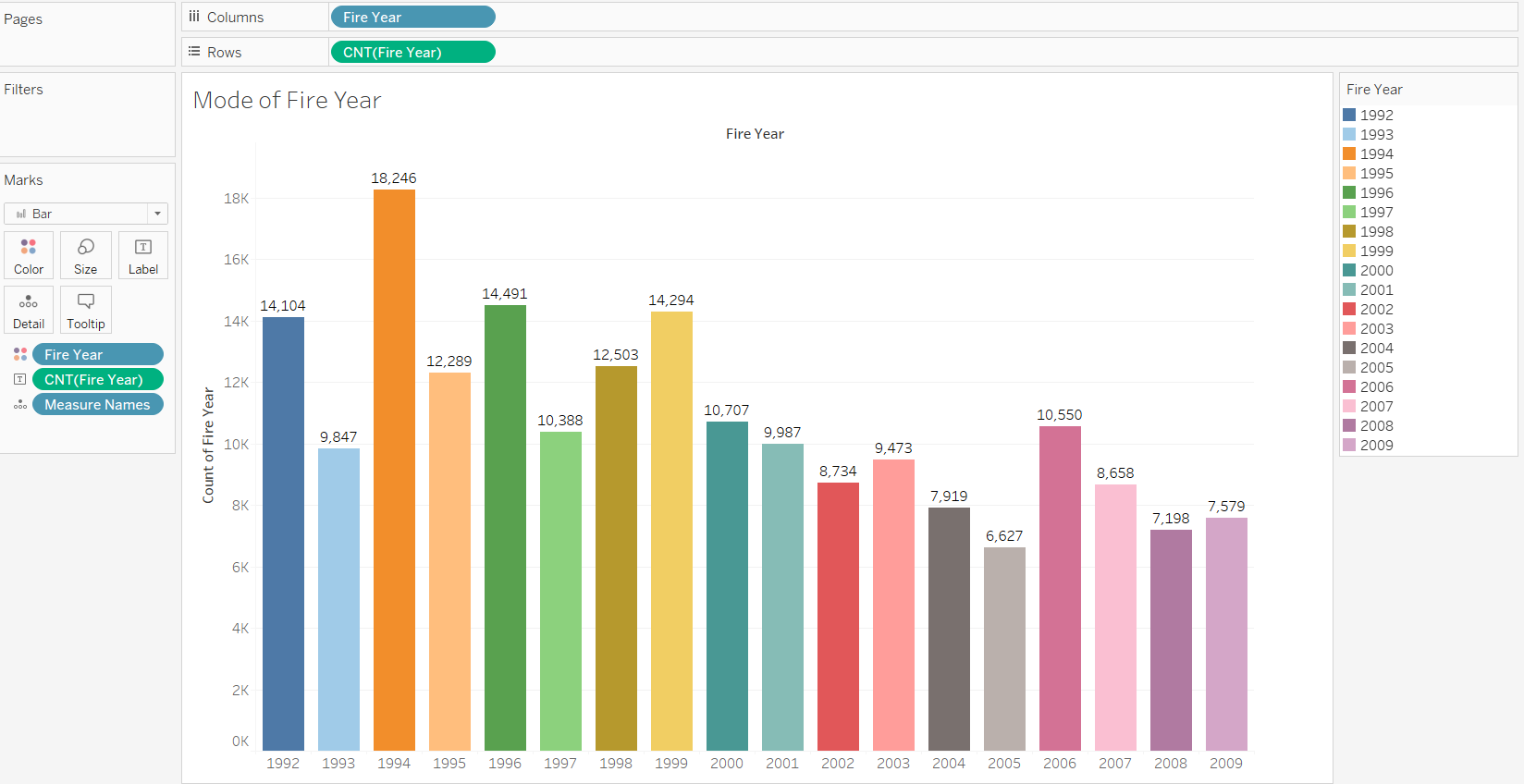


Figure 10: Mode of Fire Year

The mode of fire years occurs in 1994, with a count of 18,246. This suggests many fires occur in 1994.

**6.9 Mode of Fire Size Class (Bar Chart)**

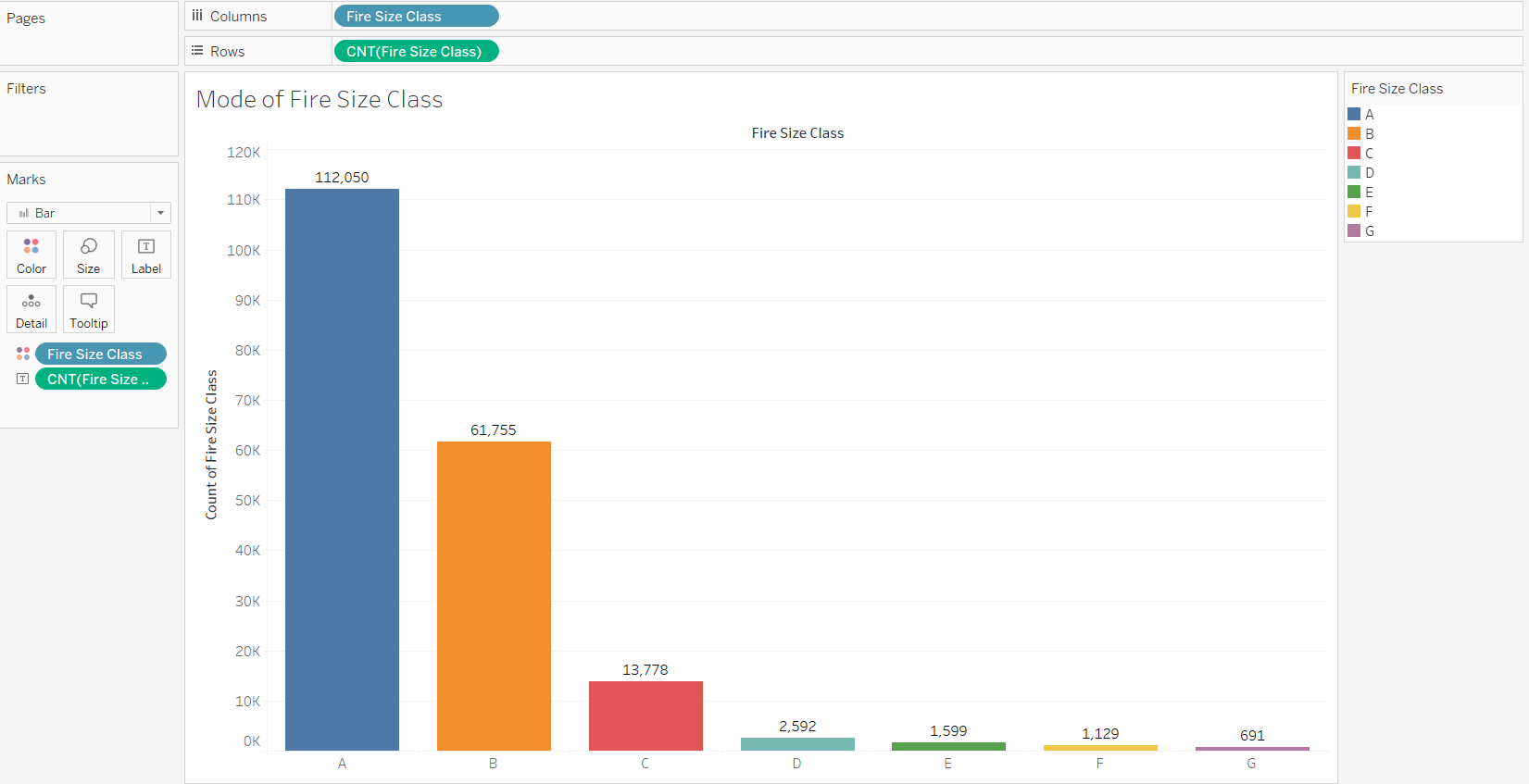


Figure 11: Mode of Fire Size Class

The mode for Fire Size Class is class A with a count of 112, 050. The next most frequent Fire Size Class is B Class, followed by C Class. As classes move from A to G, the fire size also increases. Since A, B and C are the first few levels of fire sizes, this suggests most fire sizes fall within a small fire size range.

**6.10 Percentage of Start Cause Categories (Pie Chart)**

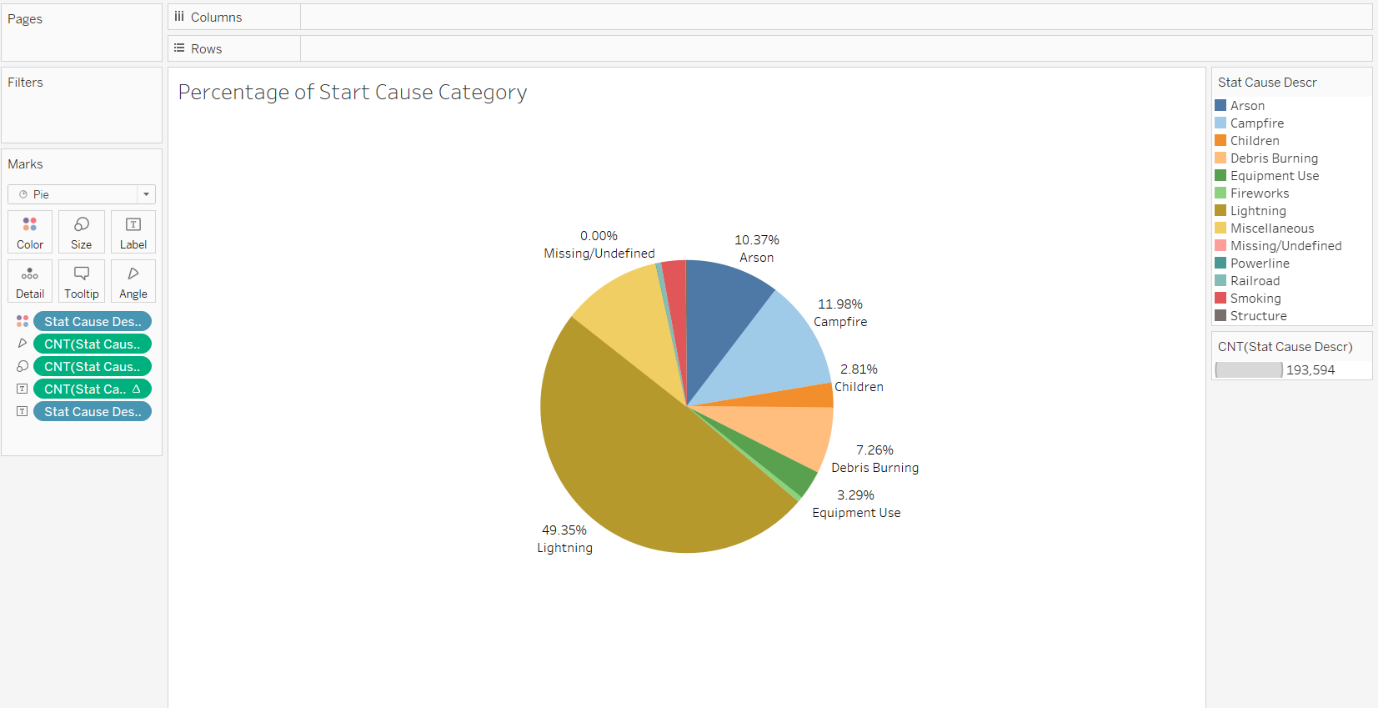


Figure 11: Percentage of Start Cause Categories

49.35% of all wildfire occurrences are caused by lightning. This means that almost half of all wildfires are caused by changes in weather.

1. **Answers to preliminary questions**
   1. **Is there a seasonal pattern in occurrences of wildfire throughout the year?**

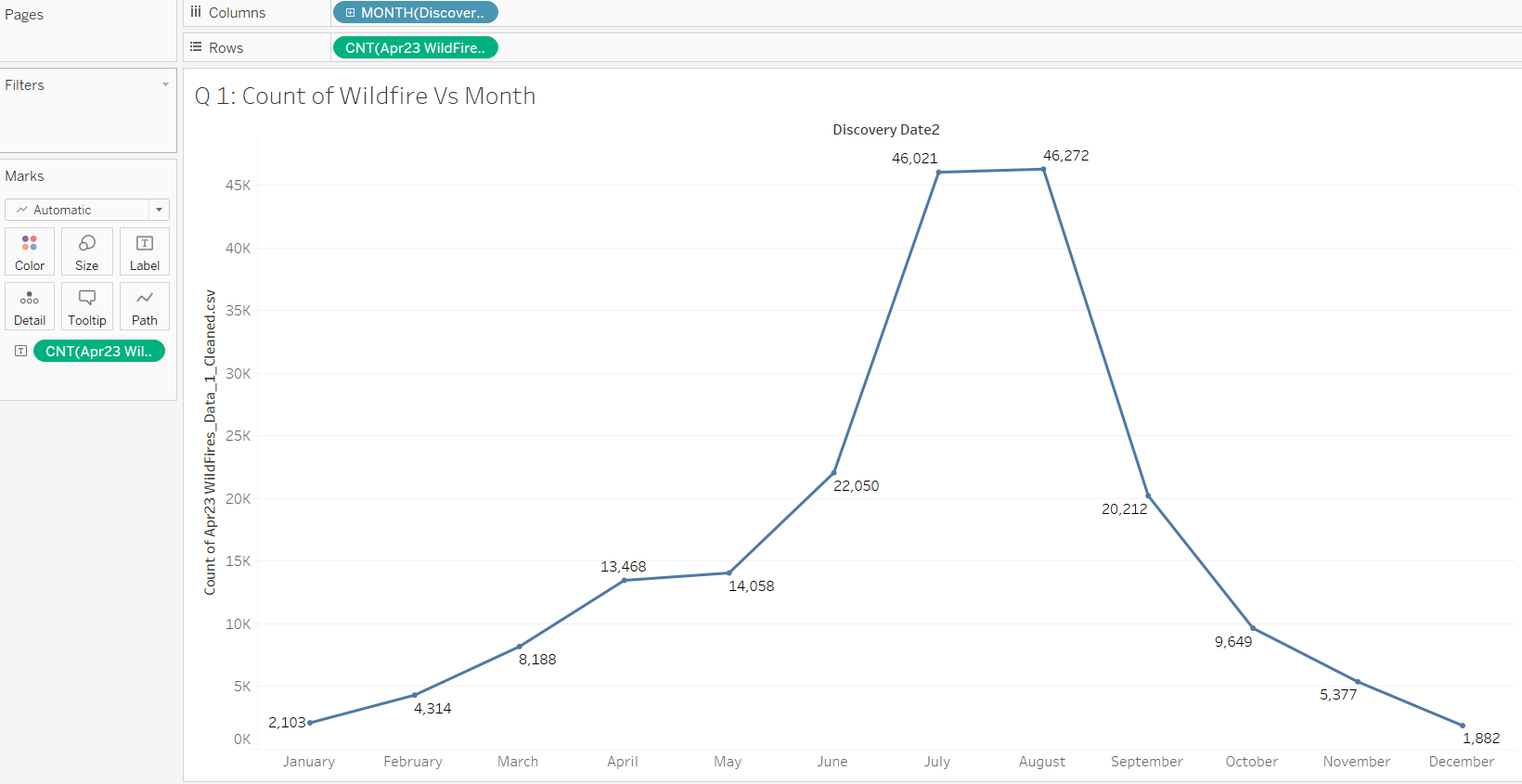
****

Figure 12: Count of Wildfire Vs Month

Throughout the year, wildfire occurrences rise from the month of January to August, then fall from August to December. From January to April, there is a steady increasing slope. From April to May, however, the slope seems to stagnate slightly before continuing with a jump in June. From June to August, there is a sharp increase in occurrences of wildfires, reaching the peak of 46272 counts in August. This could be due to the summer season, typically from June to August, where the increased heat causes the land to become hotter and drier, which are suitable conditions for wildfire occurrences. [13]

Similarly, from August to September, there is a sharp decreasing slope. Likewise, also due to the dip in temperature from changing season to autumn, causing the weather conditions to become cooler and more humid. [14] The slope becomes more gentler from the month of September to December.

* + 1. **Which states have the highest percentage of man-made compared to natural fires?**

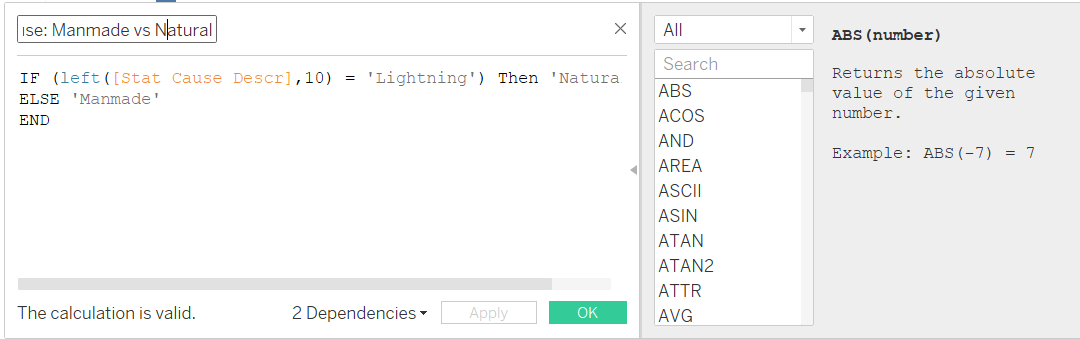
****

Figure 13: Calculated Field to Separate Start Cause Manmade and Natural

A calculated field is used to separate the categories in start cause description between manmade and natural causes. Natural causes only have lightning whereas manmade causes are the combination of the remaining start causes excluding lightning.

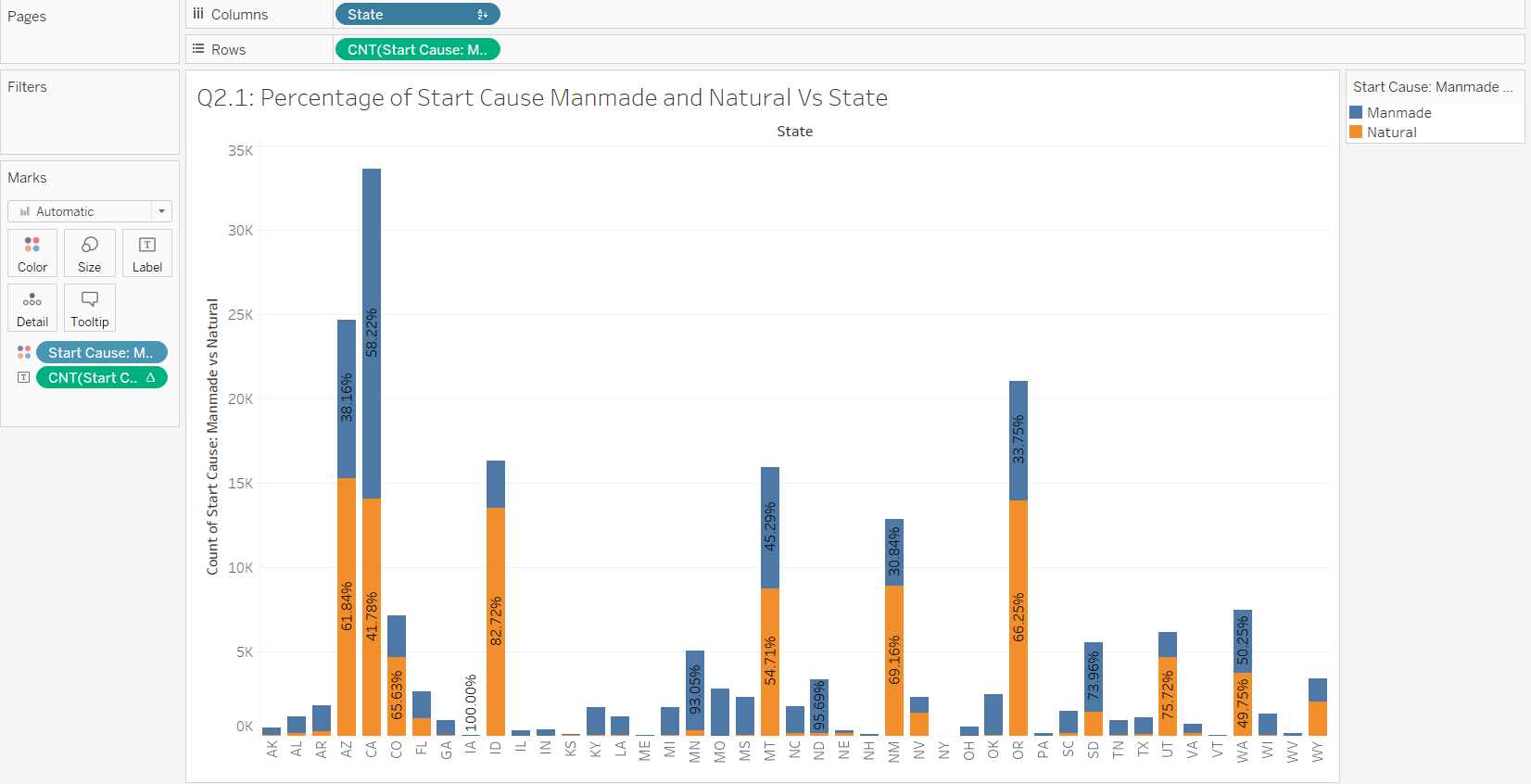
****

Figure 14: Percentage of Start Cause Manmade and Natural Vs State

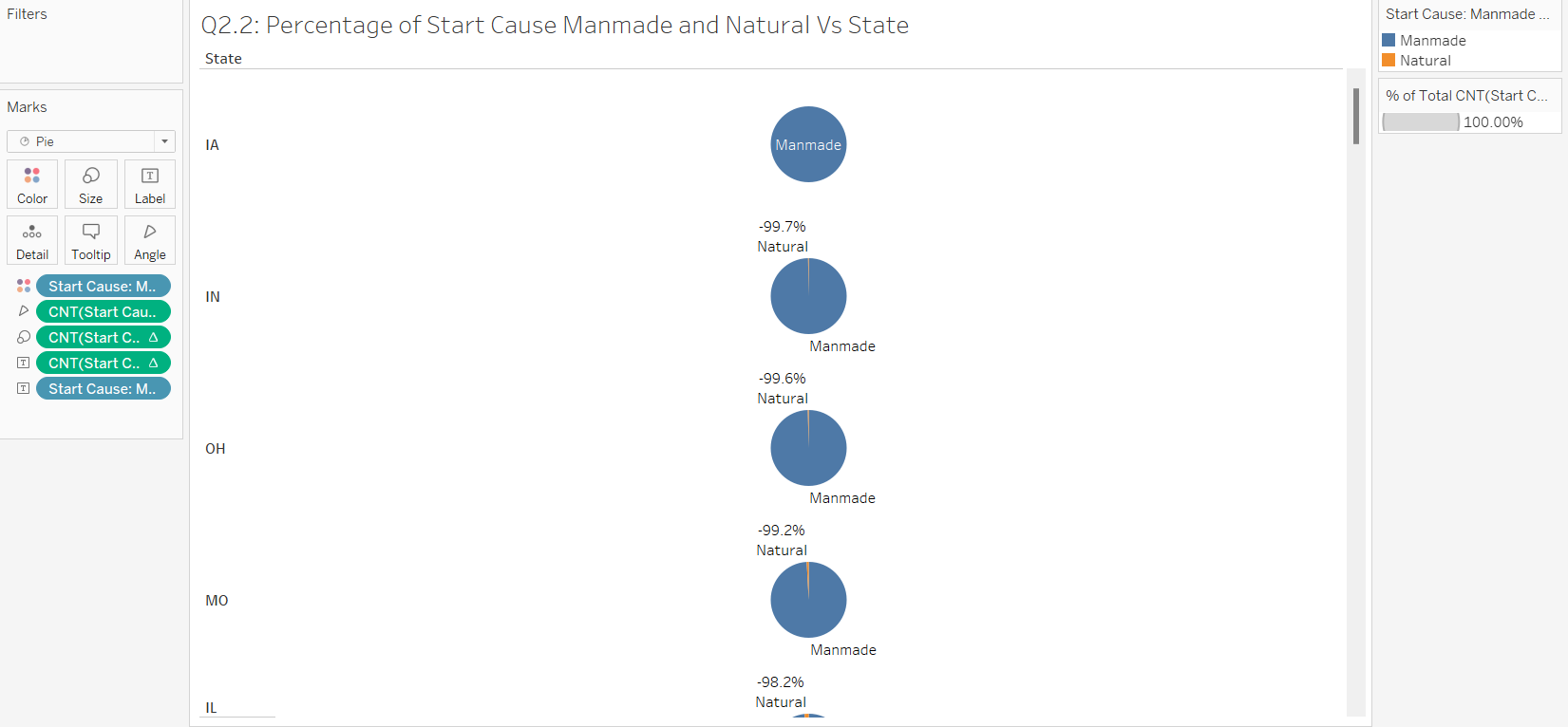
****

Figure 15: Top 3 Percentage of Start Cause Manmade and Natural Vs State

From the figure shown above, the top 3 states that have the largest percentage difference between natural fires and manmade fires were IA(Iowa), IN(Indiana) and OH(Ohio), with an overwhelming percentage of close to 100% being manmade causes.

* + 1. **In the top 3 states, which manmade causes had the highest percentage of occurrence?**

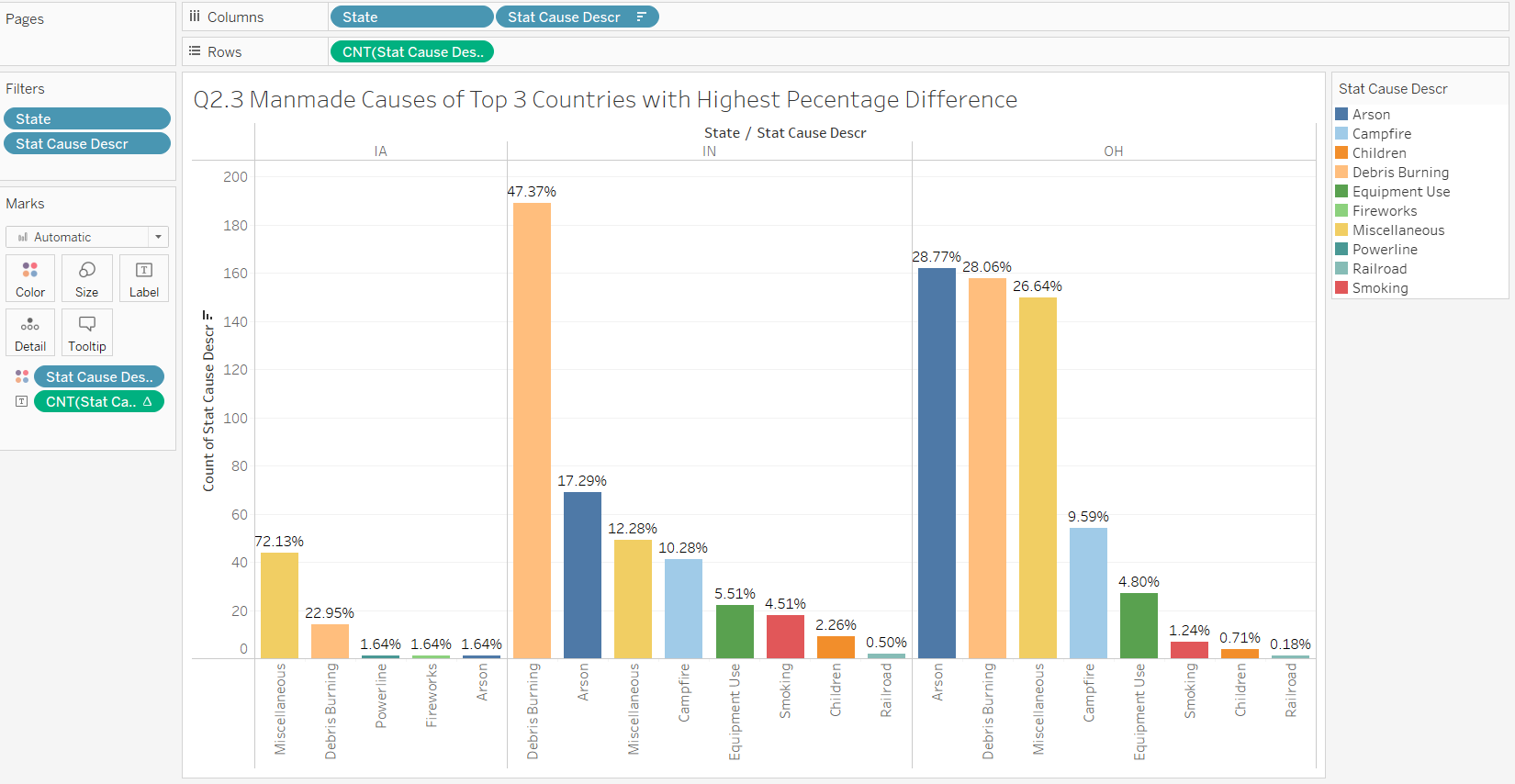


Figure 16: Manmade Causes of Top 3 Countries with Highest Percentage Difference

Further insights show that arson, debris burning and miscellaneous causes are the leading start causes to manmade fires in these countries. Most debris burning are controlled burnings necessary for clearing land, pest control and open trash burning. [15]

The larger parameters and areas of burning could have caused a higher possibility of accidental fire leakage, leading to commonly caused wildfire occurrences. An increased arson-caused fire could have been due to the rise of intentional fires caused by arsonists motivated by malice, reckless, revenge and mental health issues. [16]

* 1. **How does the average fire size change over the years?**

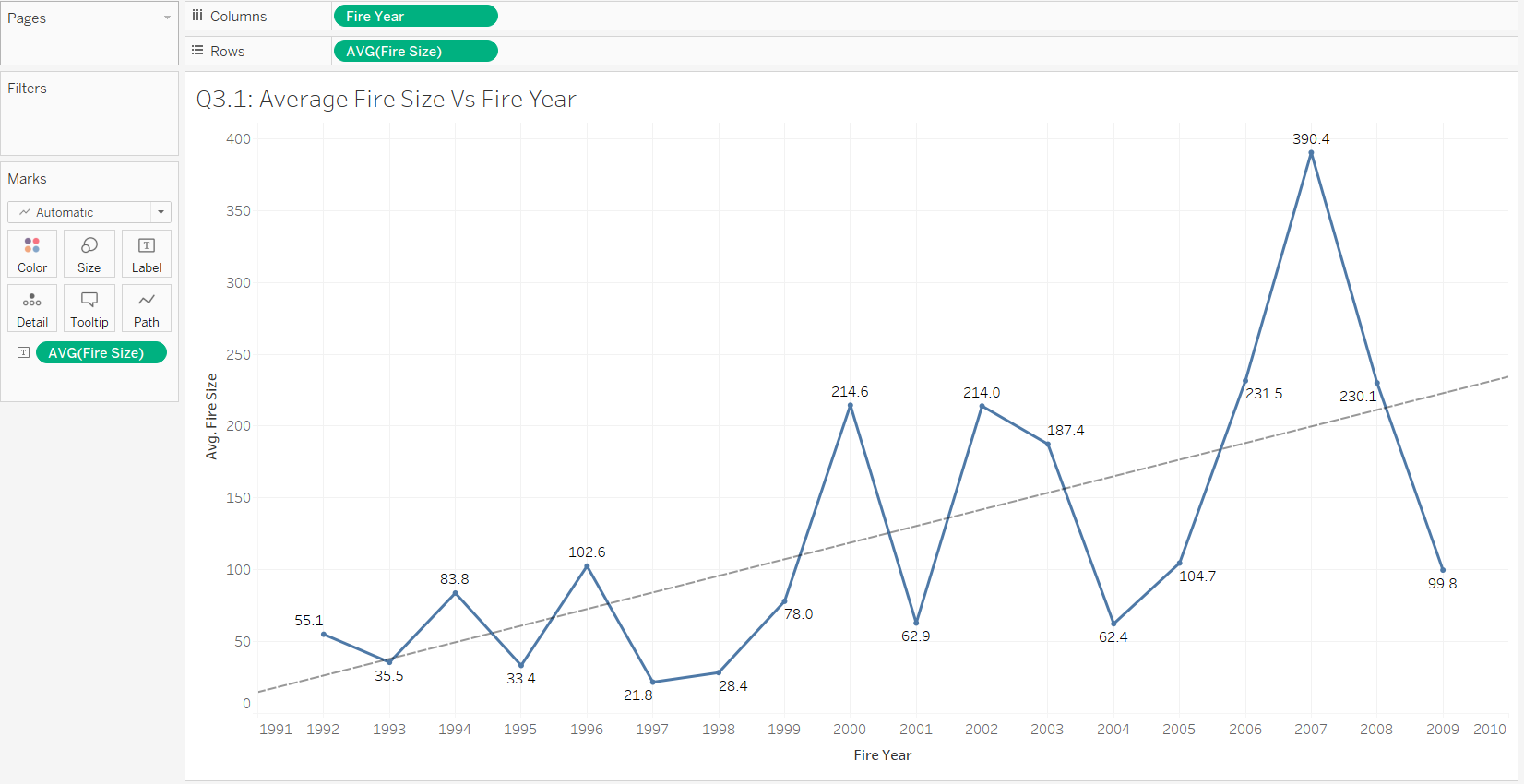
****

Figure 17: Average Fire Size Vs Fire Year

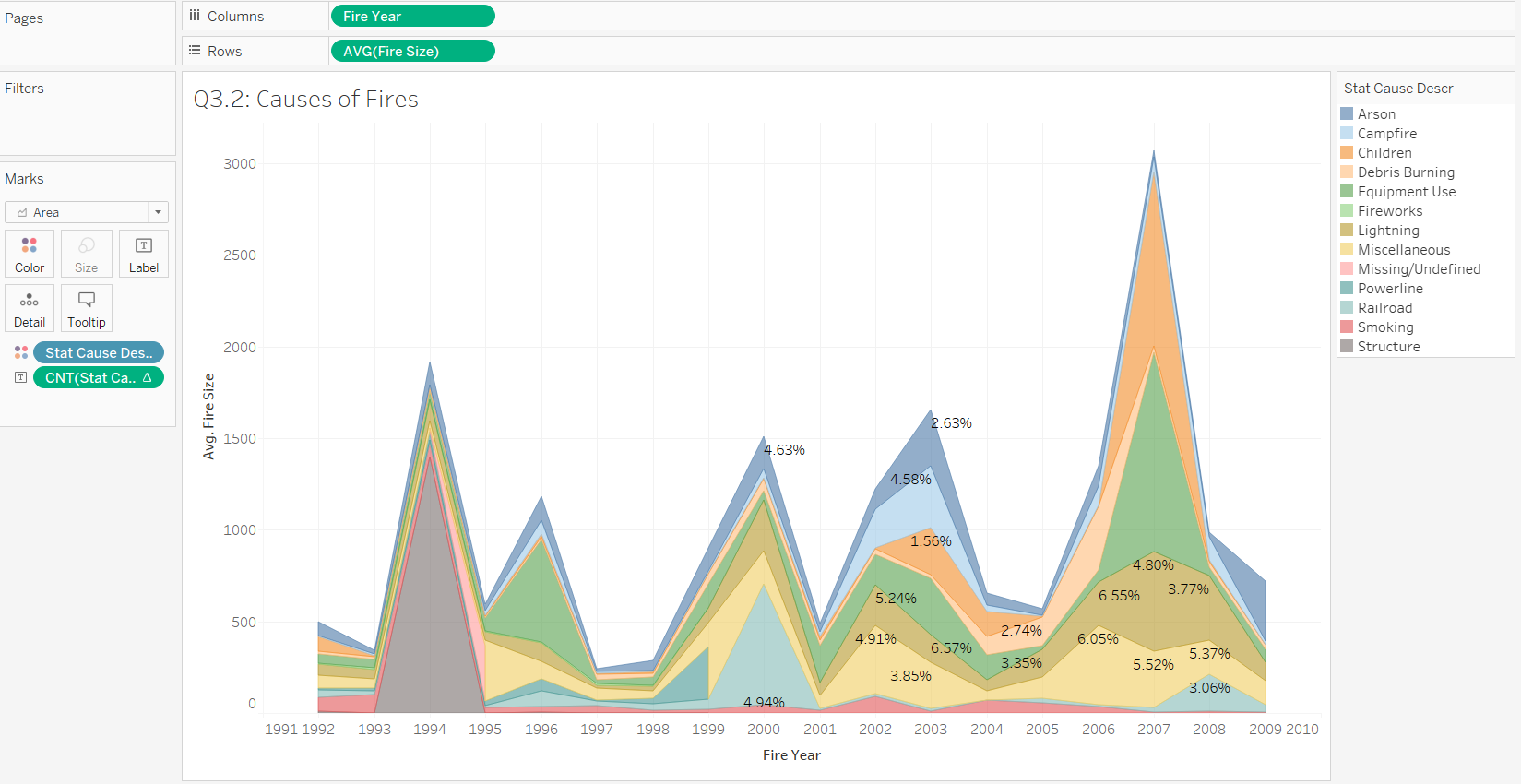
****

Figure 18: Causes of Fires

From the line graph, the trend line shows a positive gradient with a low correlation (R^2 = 0.387). The average fire size fluctuates every year, toggling from rises and dips each following year. From 2004 to 2009, there is a sharp increase in fire size, with average fire size increasing 600% from 62.4 to 390.4 (red box). Increased human activities (ignitions and fuel) and droughts are likely reasons for the rise. [17] The peak of average wildfire sizes occurred in 2007, with arsons and children contributing the most to the average fire size. The average fire size drastically drops 400% lesser from 390.4 to 99.8 (green box).

Since 2000s, climate change has been causing larger fires, increased frequencies and more widespread. These extreme fire events could have contributed to the rapid growth of fire sizes. [18]

* 1. **Is there a relationship between cause of fire and fire size?**

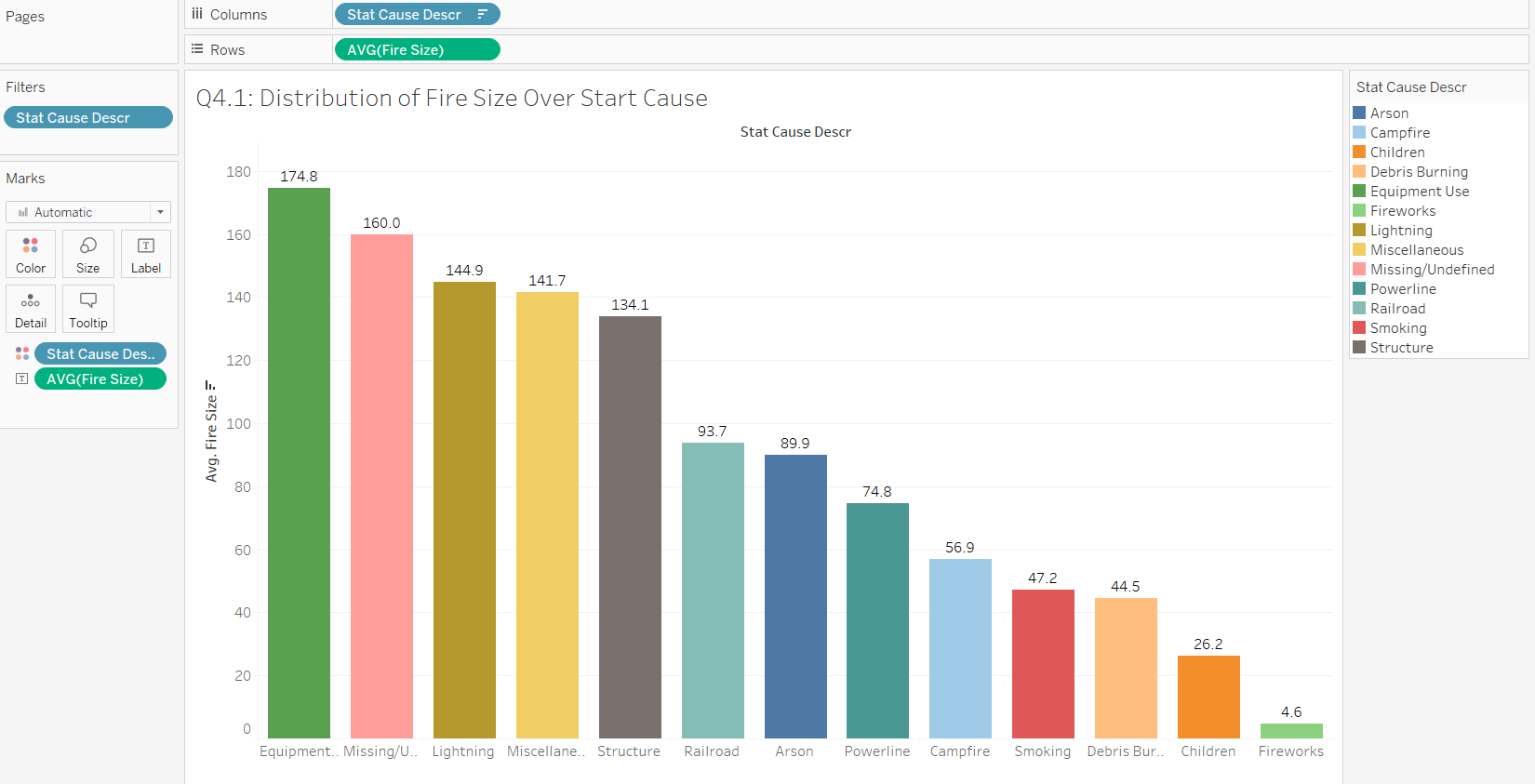
****

Figure 19: Distribution of Fire Size Over Start Cause

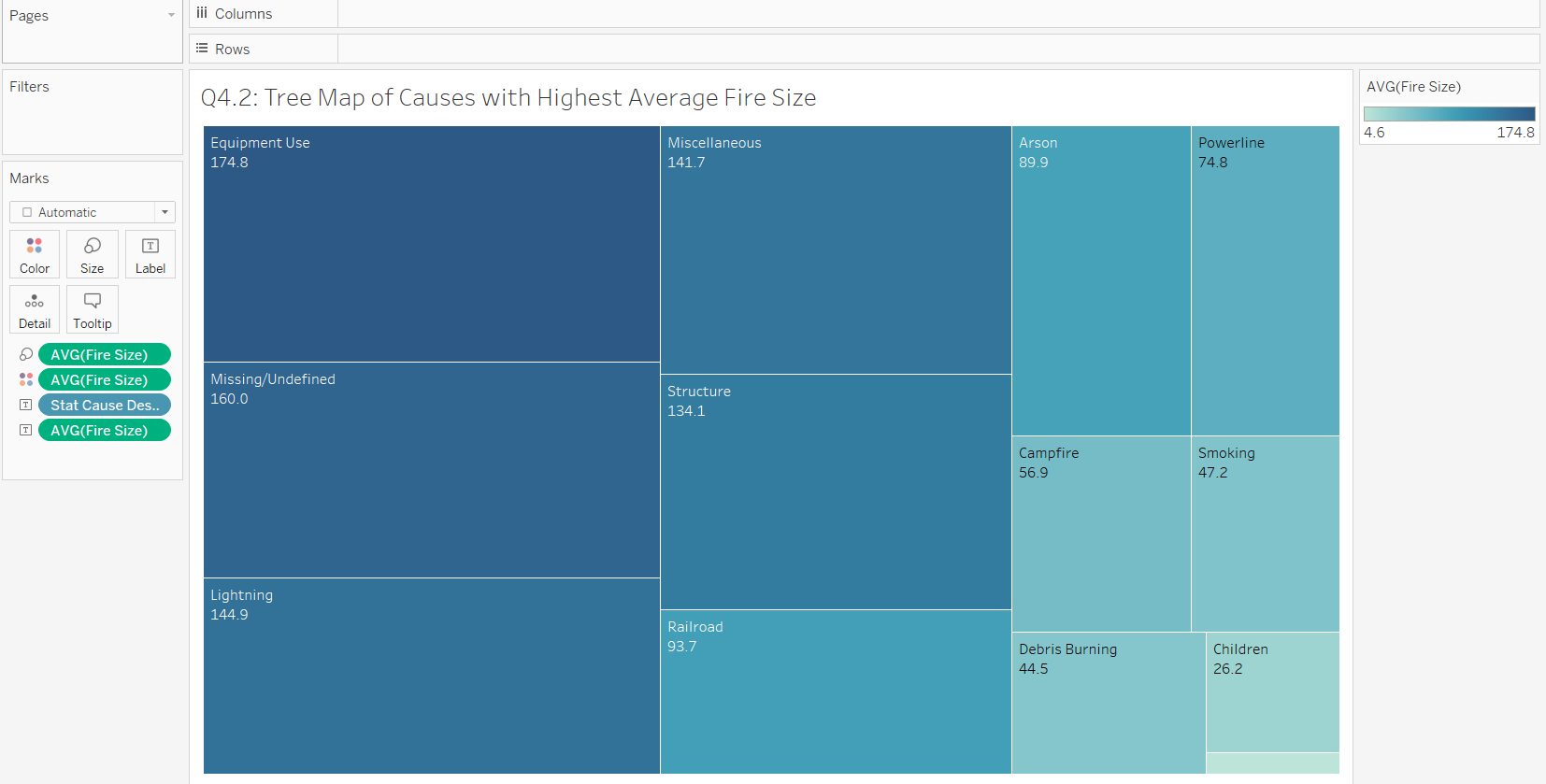


Figure 20: Tree Map of Causes with Highest Average Fire Size

The top 3 causes with the highest average fire sizes are equipment use (174.8), missing/undefined (160.0) and lightning (144.9). Equipment use commonly caused by car crashes, engine sparks and equipment failure have a large fuel source consisting of oil and flammable materials can cause a fire to spread with higher intensity. Lightning is usually accompanied by wind and storms, which can supply the fire with additional oxygen to cause it to move across a landscape at a faster rate, increasing the average fire size. [19]

* 1. **Are there any trends in the cause of wildfires over the years?**

****

Figure 21: Cause of Wildfire Vs Year

All causes of wildfire have a similar decreasing trend. From the line graph, lightning is the highest cause of wildfire. This could be due to the explosion caused by wood and other flammable materials when exposed to the high temperature of a lightning strike. [20]

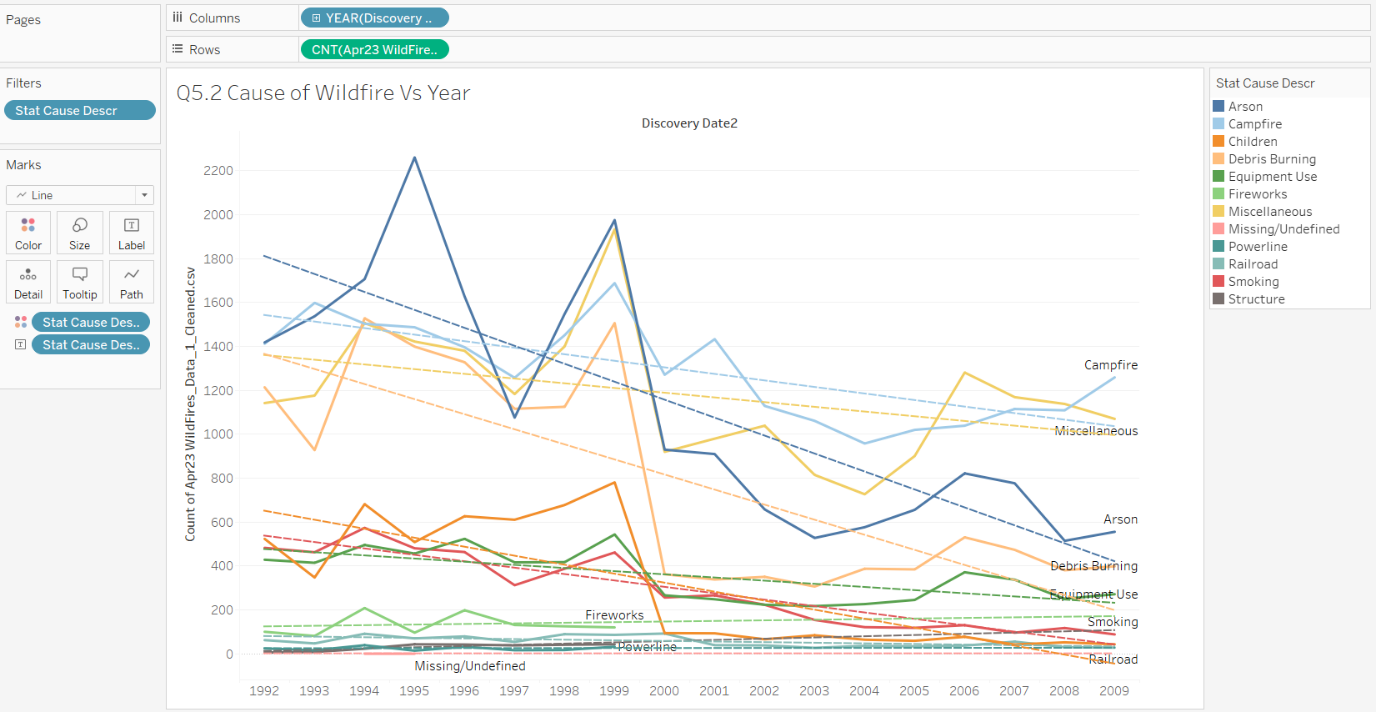


Figure 22: Close Up on Cause of Wildfire Vs Year

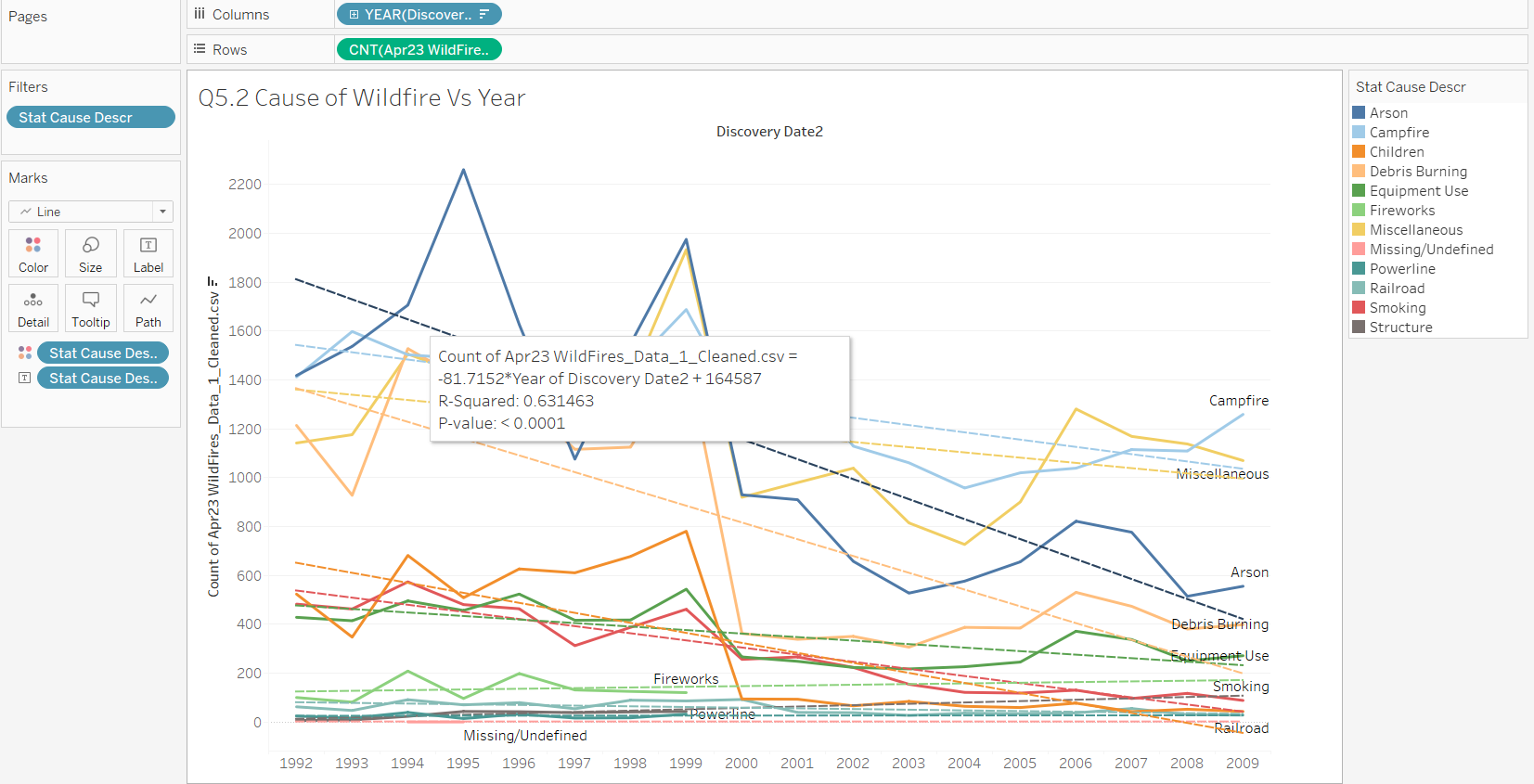


Figure 23: Close Up on Cause of Wildfire Vs Year with Correlation Values

Zooming in on the other start causes, the next most significant decline in the count of fires over the years is arson (R^2 = 0.632). Since the peak of arson fires (1995) to the lowest dip (2008), efforts such as spreading awareness and implementing regulations to prevent wildfires could have been implemented by the federal government to reduce arson.

* + 1. **Are there any hotspots of wildfires based on geographical coordinates?**

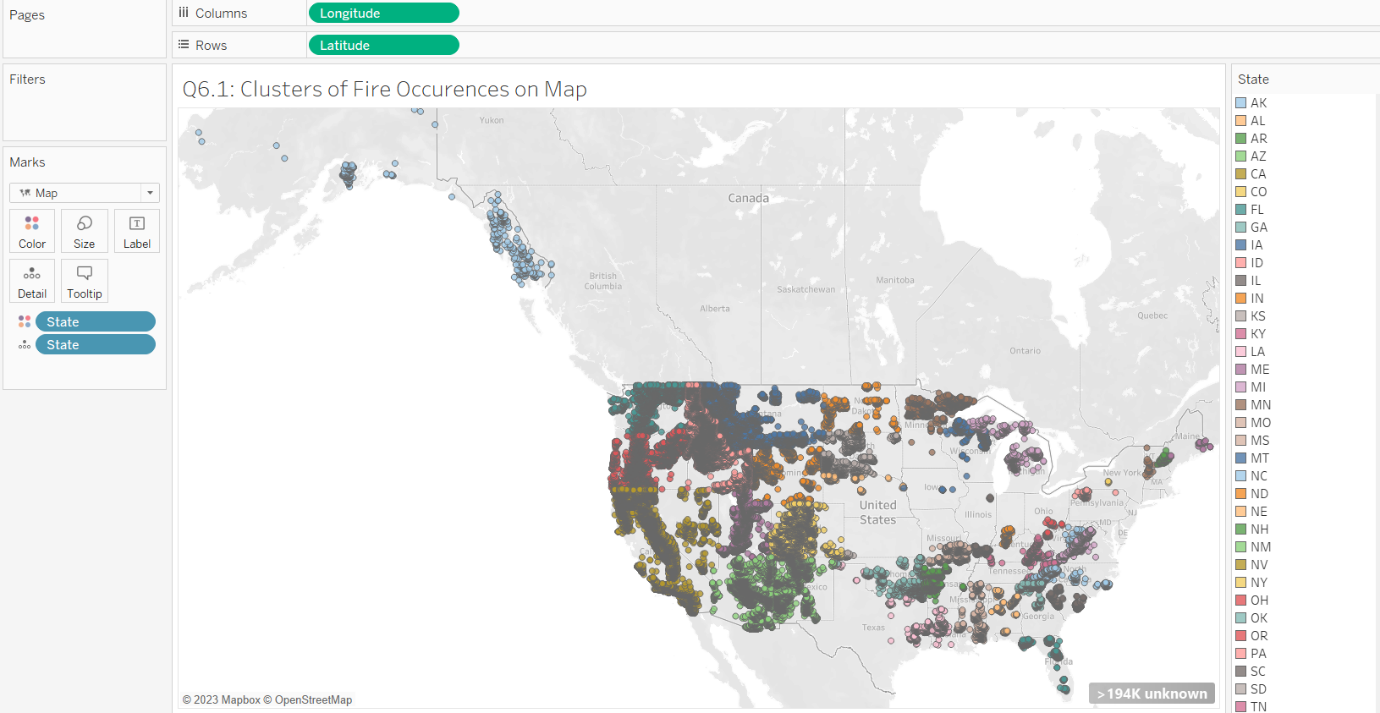
****

Figure 24: Clusters of Fire Occurrences on Map

Most wildfires occur at southeast of US. The countries bounded by the red circle are mostly forested areas that have leaves and shrubs to provide fuel for ignition of fires. [21]

* + 1. **Where are the top 3 countries with highest wildfire counts located on the map?**

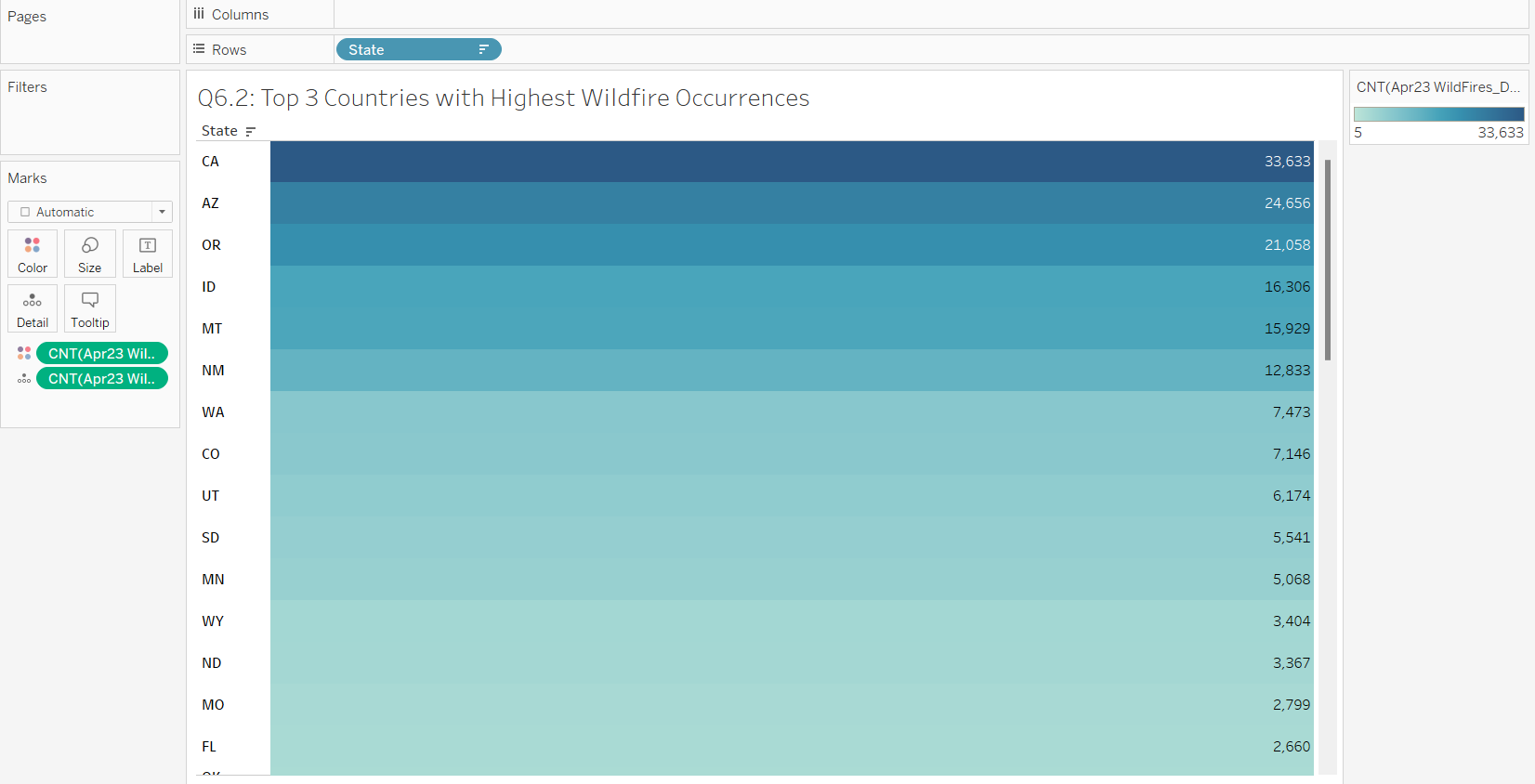


Figure 25: Top 3 Countries with Highest Wildfire Occurrences

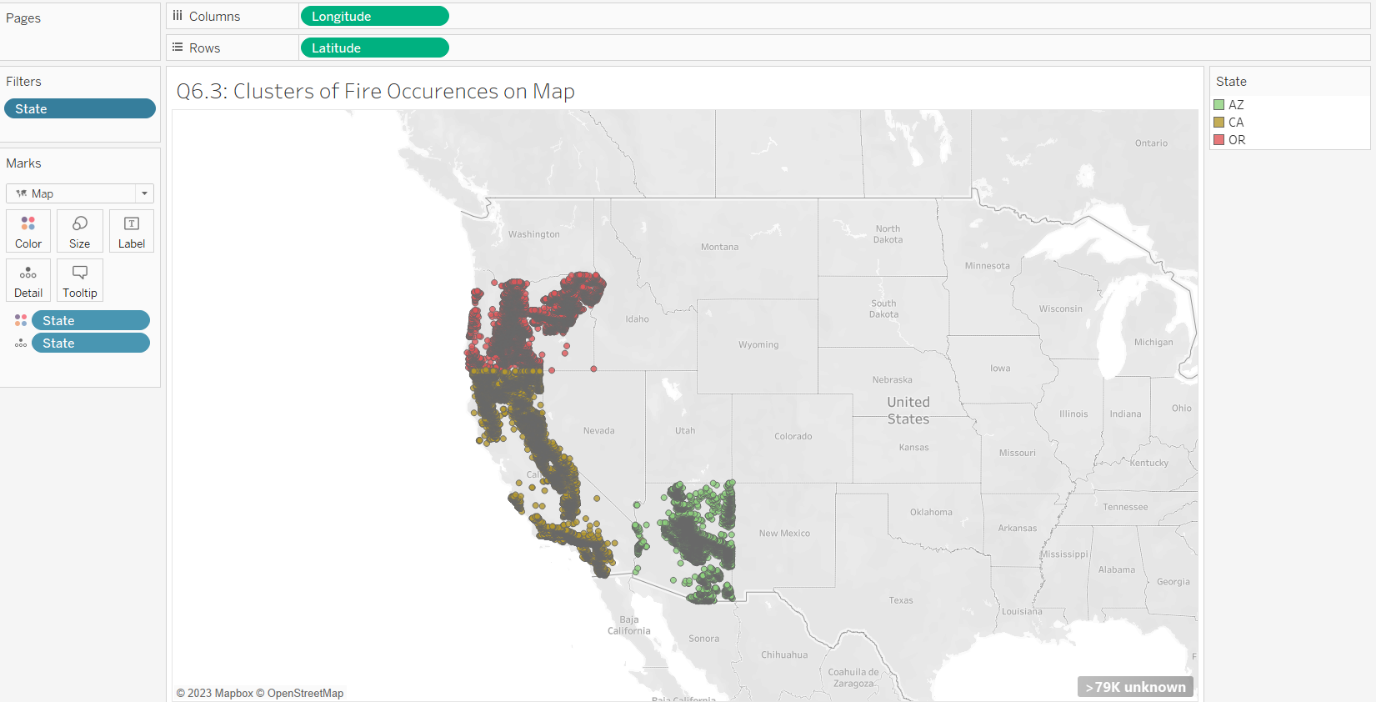


Figure 26: Top 3 Countries Clusters on Map

The top 3 countries with the highest wildfire count comes from CA(California), AZ(Arizona) and OR(Oregon), which were also bounded within the first red circle in the first figure.

* + 1. **How does frequency of wildfires differ between each reporting agency?**

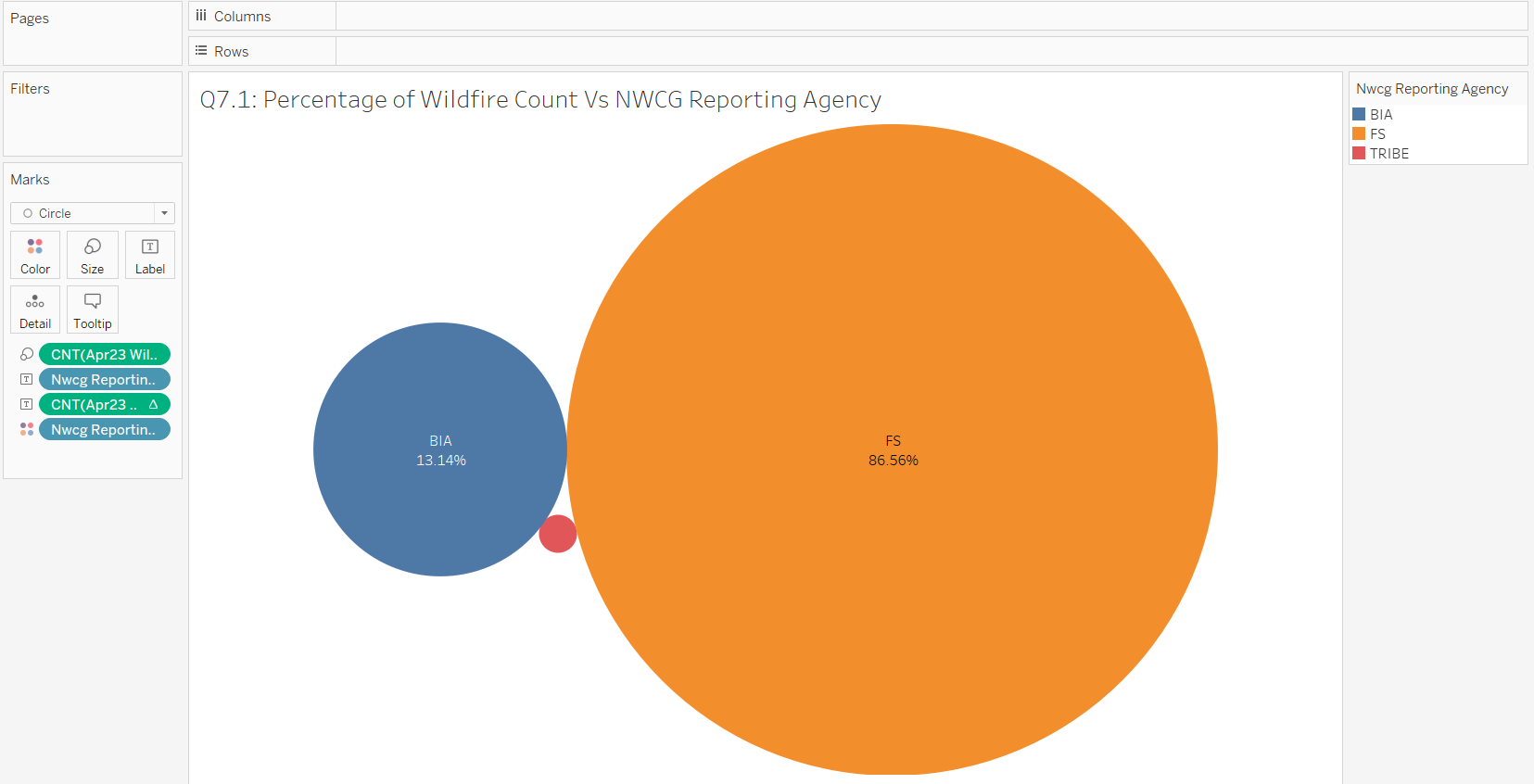
****

Figure 27: Percentage of Wildfire Count Vs NWCG Reporting Agency

FS (Forest Service) has the highest percentage of wildfire occurrences (86.56%) amongst the three reporting agencies, followed by BIA (Bureau of Indian Affairs) (12.14%) and Tribe (1.3%).

The FS is a federal agency charged with maintaining, mapping and protecting forests in US. Other than forest management, the FS also conducts forest research. These activities may need prescribed fires to manage the ecosystem and recycle nutrients back to the soil. [22] Since the FS involve many states and government affairs to manage fires, most wildfires are reported by the FS.

The BIA is also a federal agency that helps to facilitate support for tribal people and governments. There are lesser communities that the BIA need to communicate to, leading to a lower percentage in report counts. Since there are also significantly lesser tribes in US, Tribe agency report has the lowest percentage of wildfire count.

* + 1. **How does frequency of wildfires differ between each Reporting Unit IDs?**

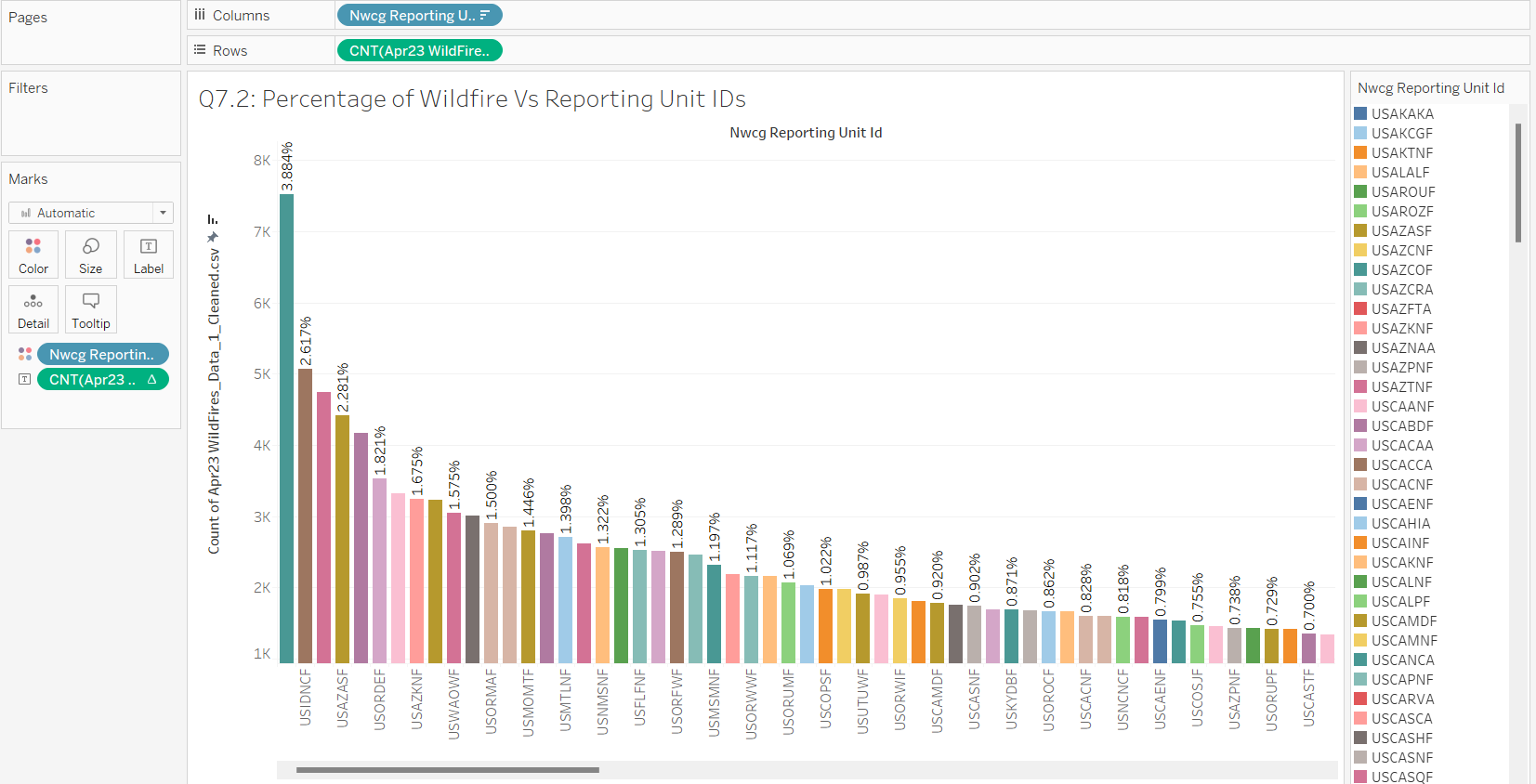


Figure 28: Percentage of Wildfire Count Vs Reporting Unit IDs

USAZCOF (3.884%) has the highest percentage of wildfire count among the spread of Reporting Agency Unit IDs, followed by USIDNCF (2.617%) and USAZTNF (2.452%). The percentage of wildfire occurrences by Reporting Unit IDs compliments the states that the units manage. This is evident when comparing with the earlier highlight table from figure 25, where the top four states with the highest wildfire occurrences are CA (California), AZ (Arizona), OR (Orlando) and ID (Idaho).

1. **Further Insights Questions (At least 3) and answers**

**8.1.1 Is there any significant impact on fire size based on the months of fire occurrences?**

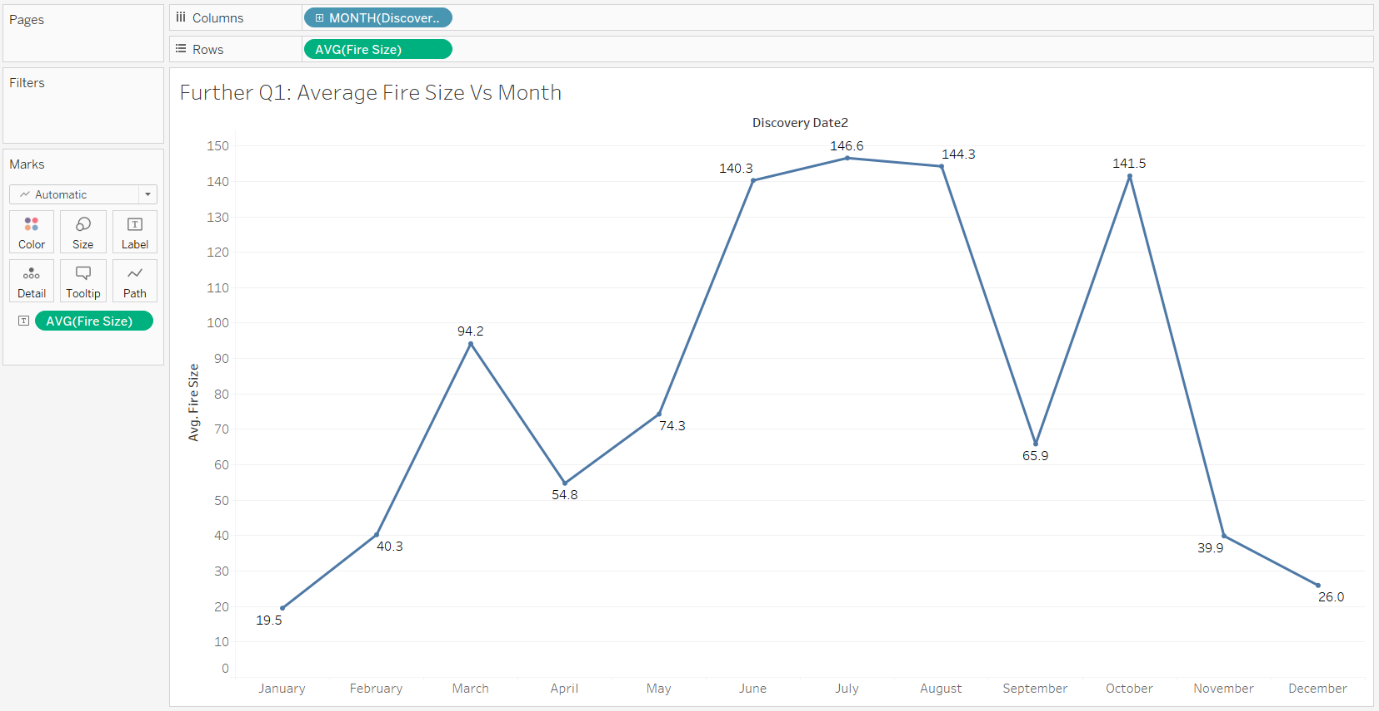


Figure 29: Average Fire Size Vs Month

From the line graph, the month of mid-May to Early August and October has the peak average fire size values among the other months. Since summer season in the US occurs in May to August, the heat and dryness of the weather could have elevated the conditions for wildfire sizes to expand more uncontrollably. Although October is in autumn, a reason for the spike could be exposure to the effects of climate change around the world leading to a rise of more extreme wildfires.

**8.1.2 On the peak month, are there visible trends in the causes of the increase of fire sizes on the map?**

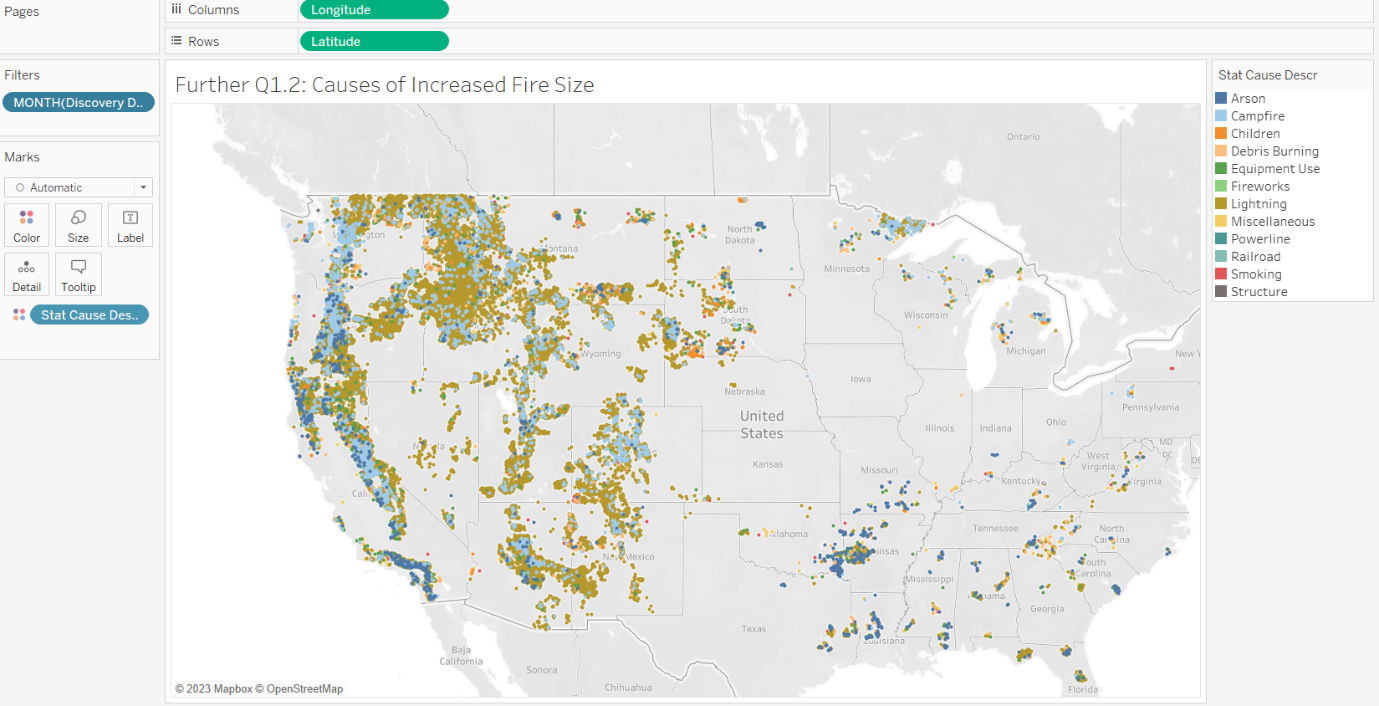
****

Figure 30: Causes of Increased Fire Size

From the map, most fires were caused by lightning, especially in the northeast of the map (red circle). Another visible cause is campfire that spreads in the east of US downwards on the map (black circle). This suggests that lightnings and campfires are one of the biggest contributors to the increase in fire size on the month of August.

**8.2 What are the most common fire size classes reported by different agencies and top 3 Reporting Unit IDs of fires occurrences for the years 1992 and 2009?**

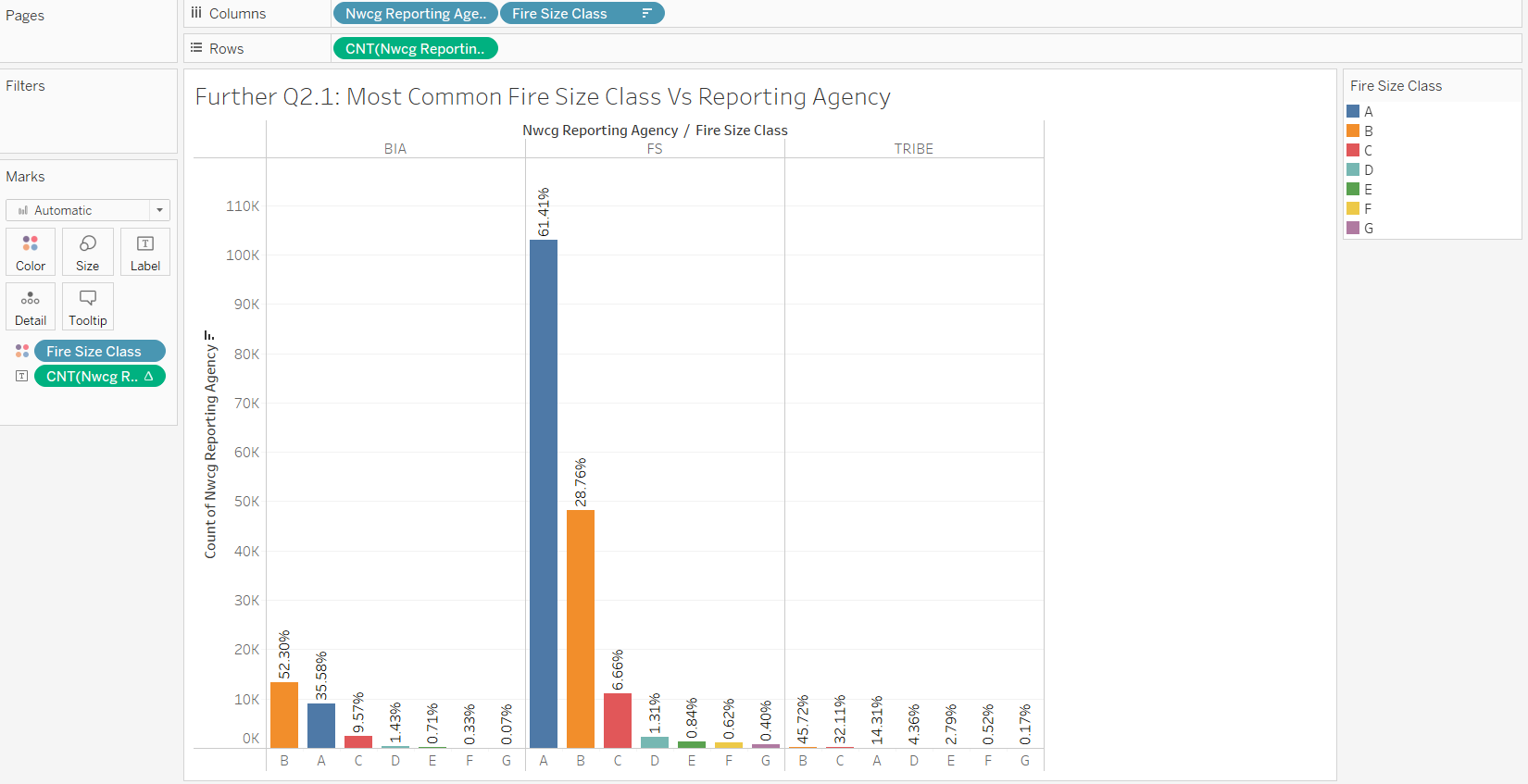


Figure 31: Most Common Fire Size Class Vs Reporting Agency

For BIA, the most common fire size class is B, followed by A and C.

For FS, the most common fire size class is A, followed by B and C.

For Tribe, the most common fire size class is B, followed by C and A.

Most fire sizes appear to fall within classes A to C which are small fires that range between 0 acres to 99 acres. This suggests the occurrences of small fires are more frequent than larger fires.



Figure 32: Most Common Fire Size Class Vs Reporting Unit IDs

Similarly, this can also be seen in the top 3 Reporting Agency Unit IDs with most occurrences of wildfires. All three Unit IDs have at least 90% of wildfire occurrences that have fire sizes fall between class A to class B.

**8.3 How does the cause of fires vary based on the state owned and private owners for the top 3 states with most fire occurrences?**

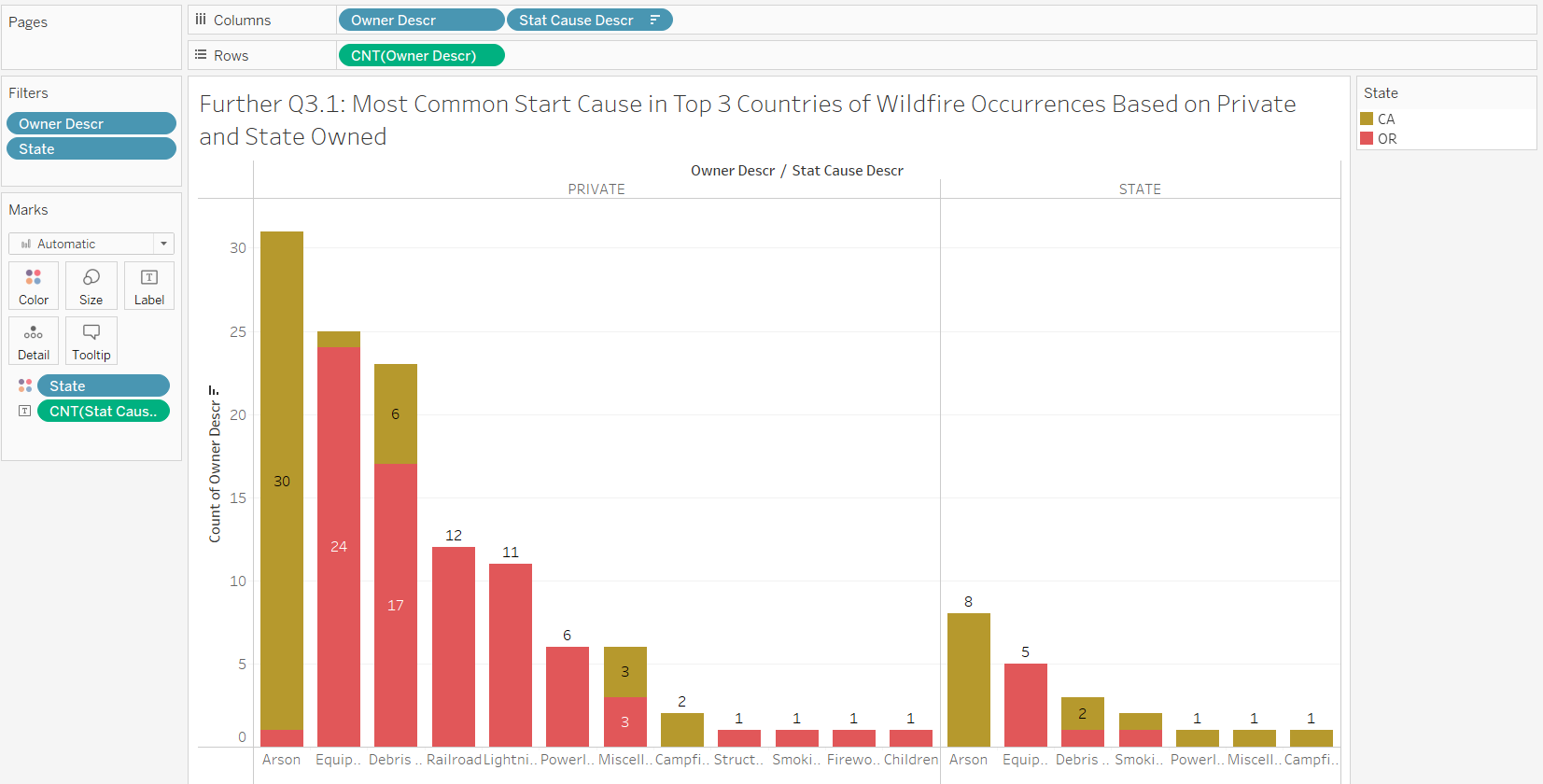


Figure 33: Most Common Start Cause in Top 3 Countries by Ownership

Only CA (California) and OR (Orlando) have state and private-owned land. Most common causes of state and private-owned land are similar, with arson having the highest count of wildfire occurrences, followed by equipment use and debris burning.

Although different landowners have their own management goals for their lands, state-owned usually determined by regulatory frameworks and public’s stewardship values, private-owned usually for industrial use, it seems that fires still occur for similar reasons.

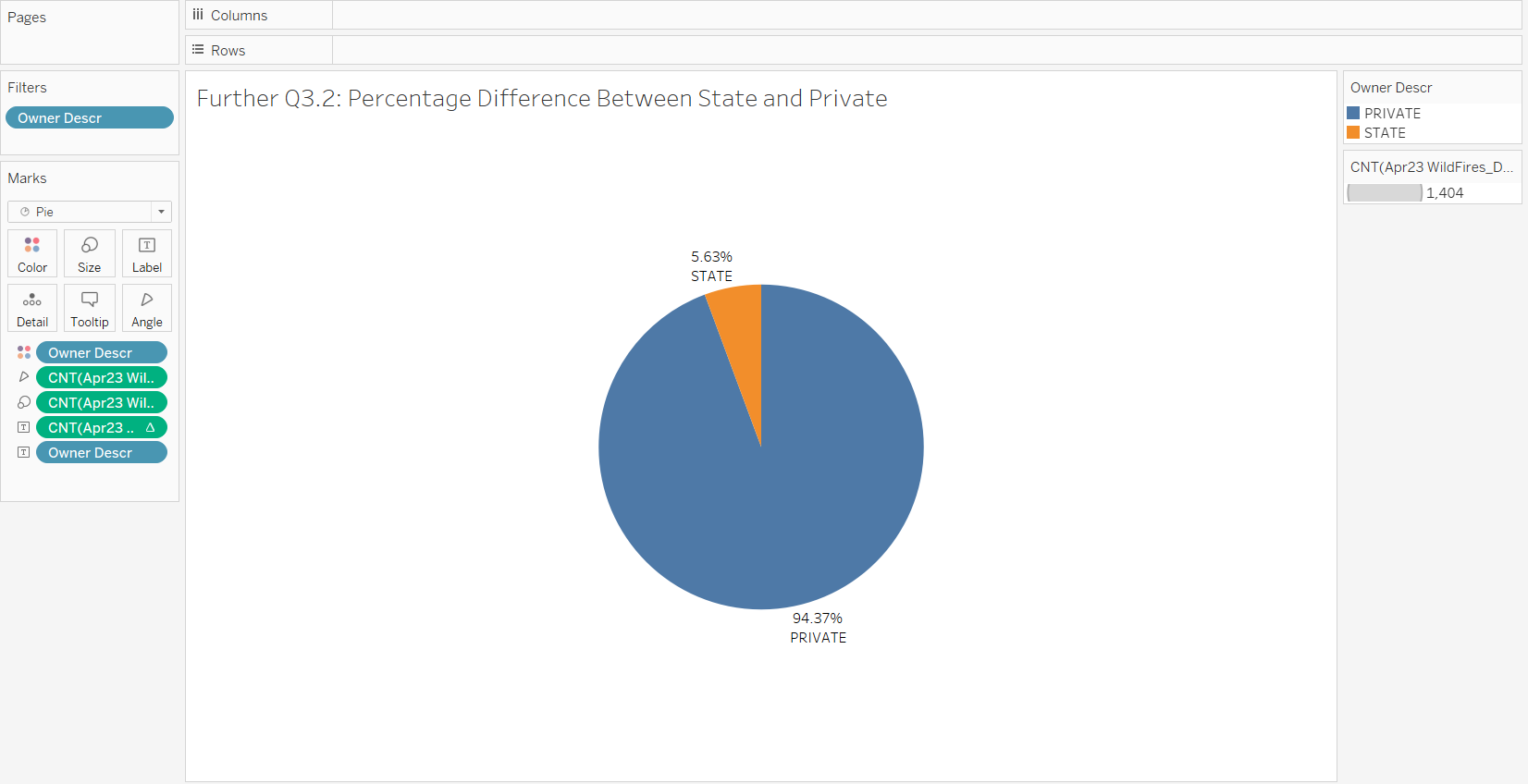


Figure 34: Percentage Difference Between State and Private Owned Land

Most fires also seem to come from private-owned lands than state-owned lands, with private-owned lands having an overwhelming percentage difference of 88.74% more than state-owned land. This could suggest state-owned lands have a better management on their land and goals, while private-owned lands require better land management policies.

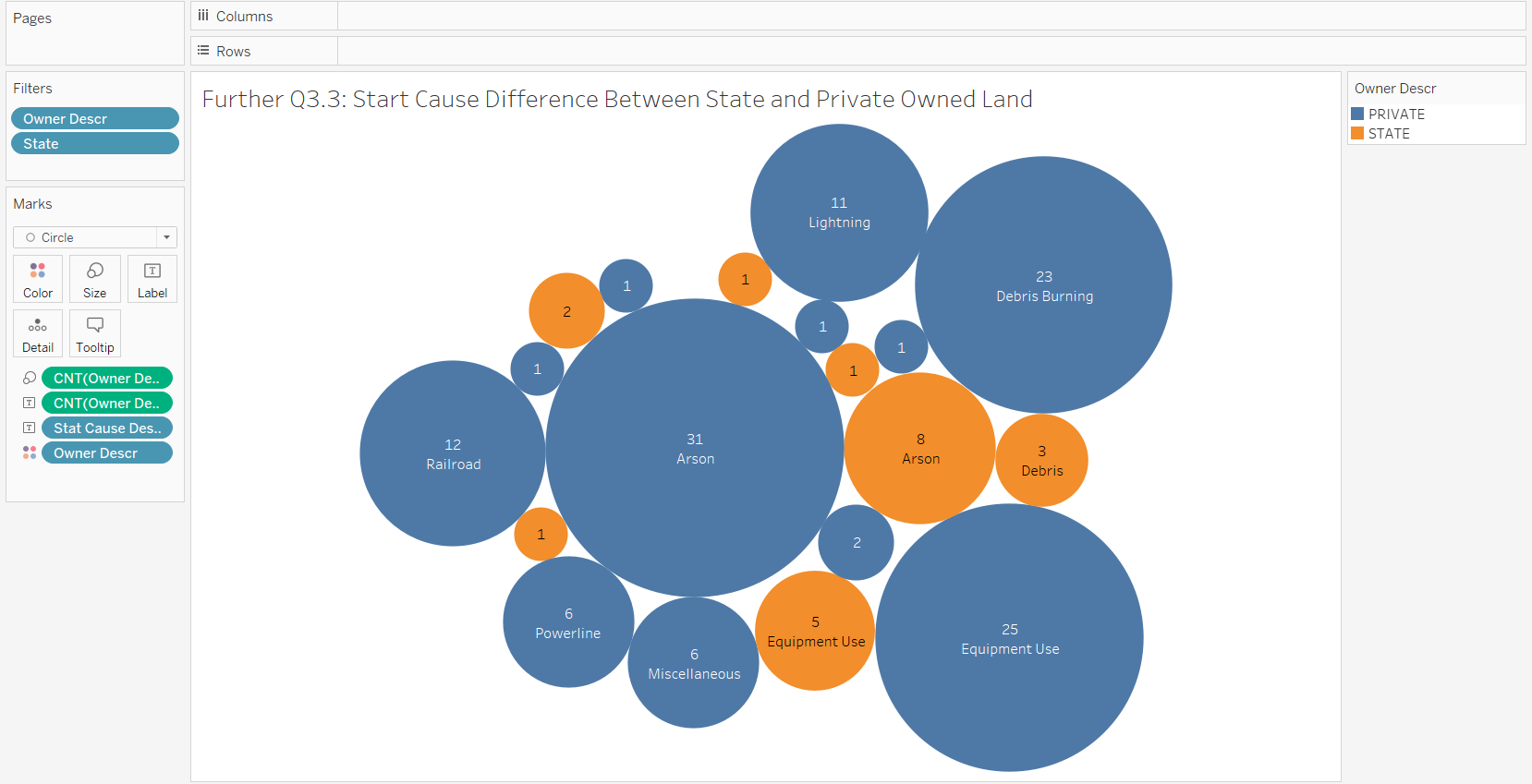
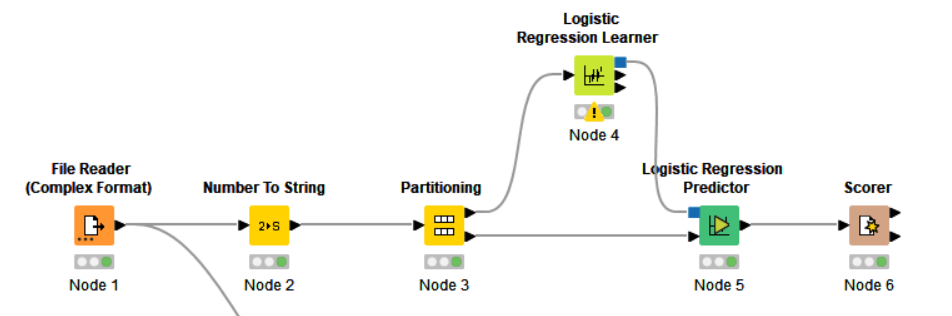


Figure 35: Start Cause Difference Between State and Private Owned Land

Private-owned land in CA (California) and OR(Orlando) have most wildfires caused by arson (31). State-owned land in these states also has most wildfires caused by arson (8).

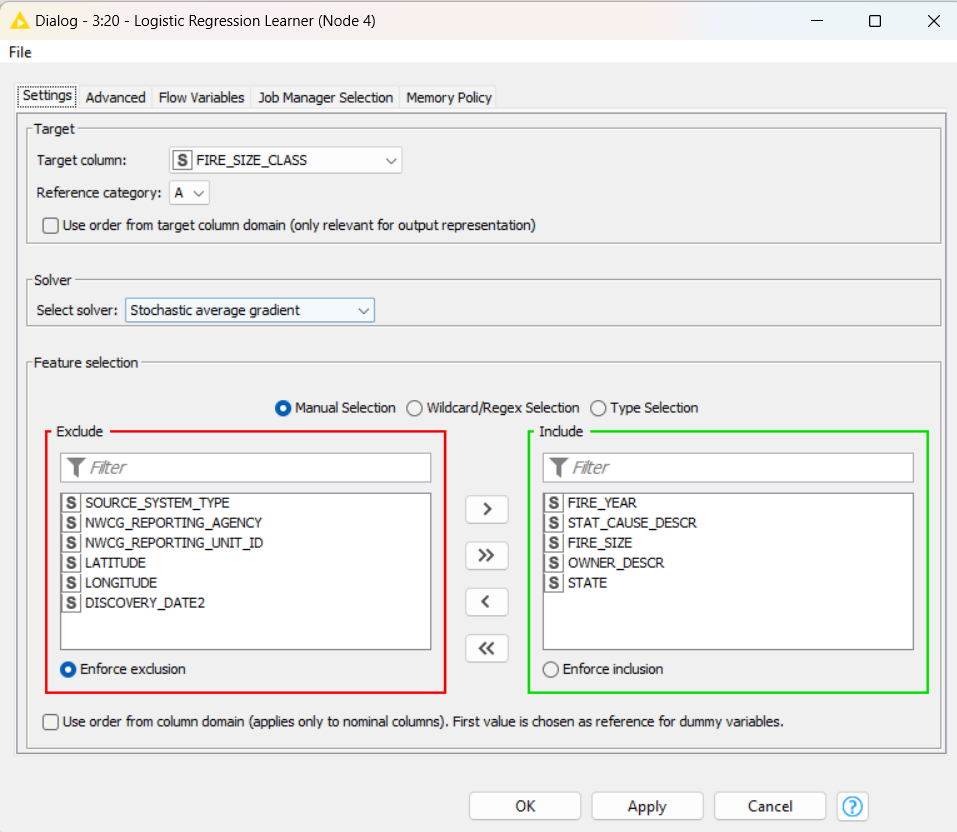
1. **Data Modeling** 
   1. **Logistic Regression Modelling**

The target for prediction is “Fire Size Class”. Since the outcome is a categorical data type, the logistic regression learner is used to generate the model. The reason why Fire Size Class was chosen as the target for prediction is because fire usually has unpredictable behavior. By predicting accurate fire size class for future fires, a proper and better preparation can be taken to reduce the risk of wildfire growth, saving lives of firefighters and those involved, as well as reducing fire extinction costs.

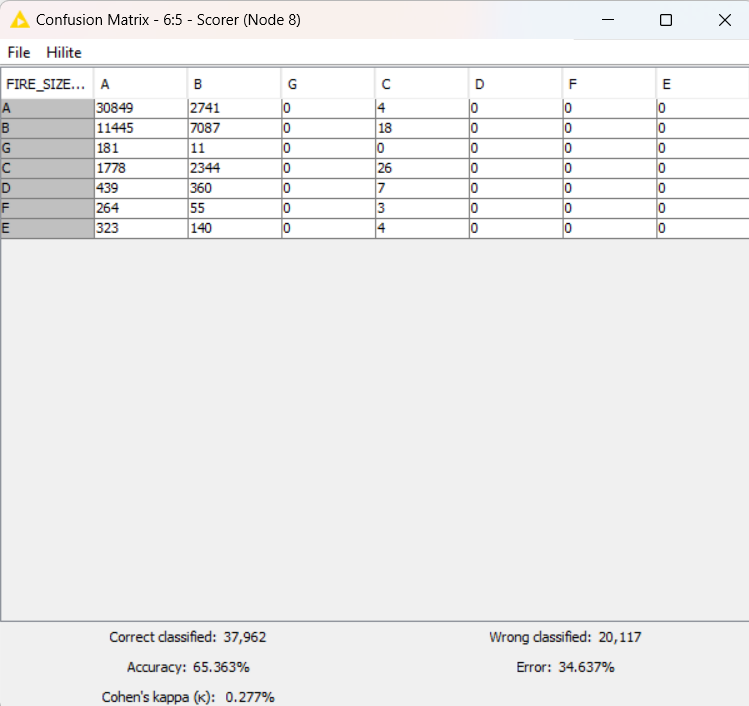


In order to generate the model, we can use:

1. Using the Number To String (Node 2), numerical data types such as Fire Year and Fire Size are converted to string data type for the logistic regression model.
2. Using Partitioning (Node 3), data is split 70% for a training set and 30% for a test set. The training set will be used to determine the model (algorithm), while the test set is used to validate the model.
3. Using the Logistic Regression Learner (Node 4), we can select our variables to be tested with our target variable. The model (algorithm) is determined by using the training set. For this dataset, we will be using string data to predict our target variable, fire size class. All other selections are default.



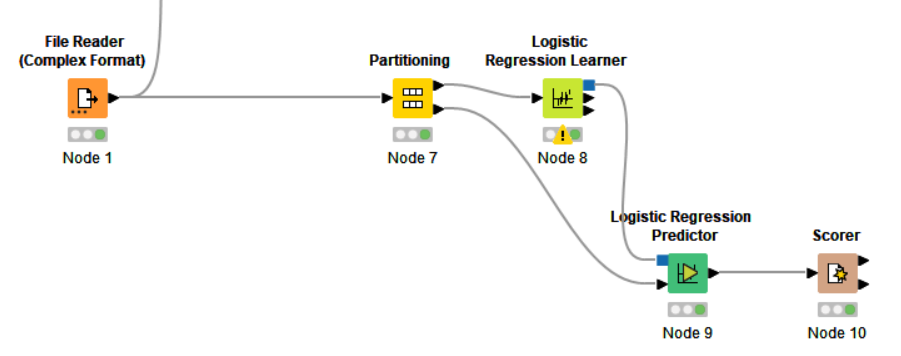
1. Using the Logistic Regression Predictor (Node 5 and 6), our model is applied on the test set. The target variable (fire size class) is predicted.
2. Using the Scorer (Node 7 and 8), the result is tabulated into a confusion matrix.

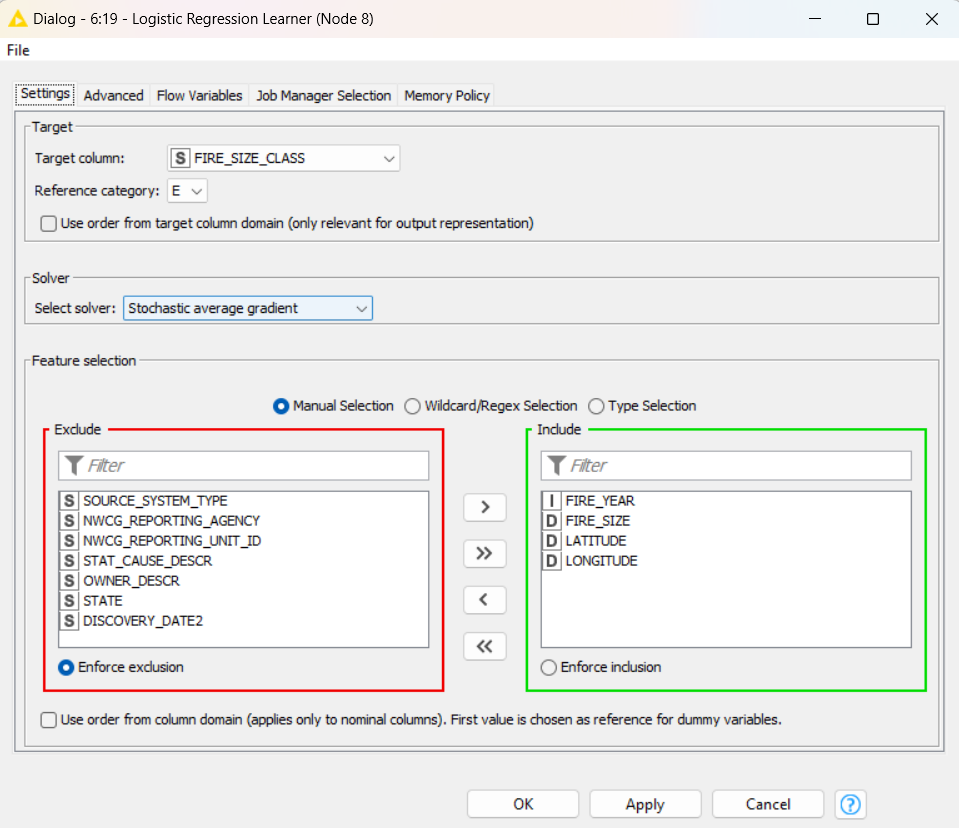


This is the result of the confusion matrix with all string input data. The matrix shows that:

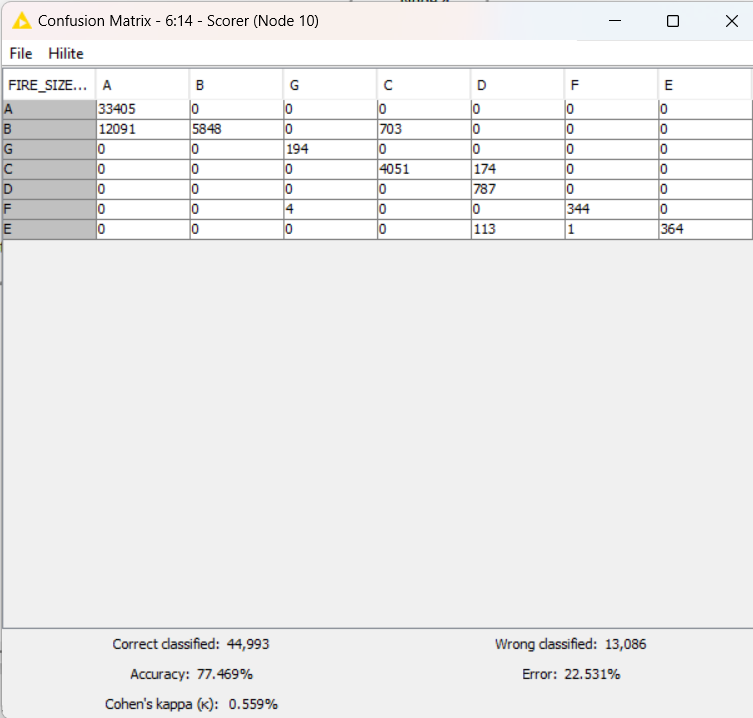
* The Correct classified data is 37962/20117. Since overall classification of data is mostly correct, the prediction will be more accurate.
* The accuracy of the model is moderately good at 65.363%. As percentage accuracy increases, the model will be more correct in predicting the data of observed values. Thus, this model has moderately good accuracy in predicting data values. With a relatively fair error of 34.637%, this model has a moderate percentage of error and will less likely make mistakes when predicting the data for the observed values.
* Cohen’s kappa (k) value is 0.277%, suggesting that there is a fair inter-rater reliability as the higher the kappa value, the higher the possibility of the agreement occurring by chance.

* + 1. **Logistic Regression Modelling Improvements**





For this workflow, only numerical data types are used for the prediction. This is configured in the Logistic Regression Learner (Node 8), including only numerical data types for the model (algorithm).



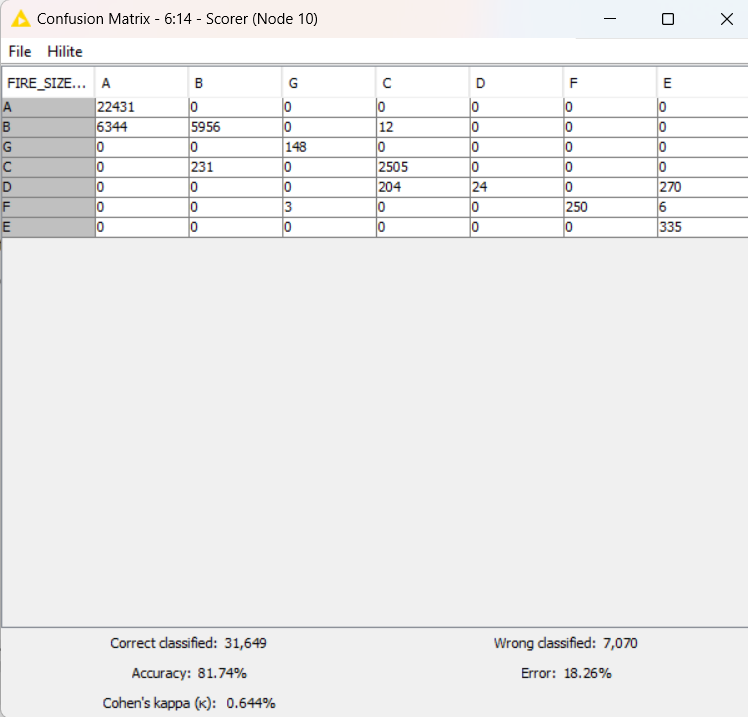
The confusion matrix displays the results. From the results, we can see that:

* There are more correct classified values to wrong classified of 44993 to 13086. This suggests that the logistic regression learner is better at predicting the targeted value correctly using numerical data types than string data types.
* Both percentage of accuracy and error increases and decreases respectively by 12%. Since accuracy increases, this means that the predicted value will have a higher probability of occurring. The decreased error percentage also helps to lower the error rate of the predicted outcome.
* Cohen’s kappa increased by 0.282% from 0.277% to 0.559%, which suggests that the inter-rater reliability is better and more reliable.
  + 1. **Best Accuracy Model**

**A screenshot of a computer

Description automatically generated**

Using the same workflow, Partitioning (Node 7) is changed to 80% for training set and 20% for testing set.



The results provided by this confusion matrix were the best among the three predicted matrices. This matrix shows:

* There were much more correct classified values of more than 75% than wrong classified. This will significantly improve the accuracy of the predicted target value obtained.
* The accuracy percentage is very high at 81.74%, with a very low error of 18.26%. The higher the accuracy, the model will correctly predict more data for the observed values. Thus, this model has a high accuracy in predicting data values. The model will also be more unlikely to make mistakes when predicting data for the observed values, having a low error rate in predicting data values.
* Since Cohen’s kappa value is 0.644%, there is a good inter-rater reliability. This outcome will have the highest possibility of the agreement occurring by chance among the 3 results.

**10. Conclusion**

In conclusion,

* Wildfire behavior is unpredictable as there are very weak linear correlations between statistical variables.
* Fire size classes mostly fall in class A, which is a small range fire.
* Wildfires are mostly caused by lightning, which is a weather condition. This could be due to climate change causing extreme weather events, increasing risk of wildfire occurrences.
* The summer season between May to August is the worst months for wildfire to occur. This is because the hot and dry conditions elevate the environment for wildfire to grow and spread, leading to a bigger fire size.
* Iowa, Indiana and Ohio require better law management and public awareness of the dangers and risk of wildfires. Since most causes of wildfires in the area are manmade, a better control and supervision on the activities of the people living in those state could reduce and prevent wildfire occurrences.
* There is an increasing trend in average fire size over time. This is also most likely caused by climate change affecting weather patterns and conditions, leading to increased size of fires.
* Equipment use is the most dangerous cause of fire as the average fire size is the largest.
* Efforts have been made to reduce manmade fires, proven from the significant decrease in arson-caused fires over time.
* Most fires occur in forested regions in the US, where leaves and shrubs provide fuel for increased ignition rates.
* California, Arizona and Oregan have a high fire count, most likely due to locating near forested areas.
* More resources could be given to USAZCOF, USIDNCF and USAZTNF Unit IDs as the reports of wildfire counts by these Unit IDs are quite high, hence requiring more help and staff.
* Private-owned lands need to be more aware and firmer on taking actions to prevent wildfire from occurring.
* From the logistic regression learner, the learner is better at predicted target values using numerical data than string data.
* From the logistic regression learner, the prediction for the target “Fire Size Class” with the final try will be quite accurate as the percentage of accuracy is 81.74%.

**11. Reflection**

Personally, I found completing this report challenging. Firstly, managing my time was difficult because there were too many ongoing projects that were due on a similar date. I had to learn how to prioritize doing which subjects and parts specifically, scheduling my time so that I could be most productive. Learning to improve myself and time management skills were something I had learnt and overcame during this process. Eventually, I still managed to keep track and finish the report by the target dates I had set for myself.

Secondly, I had found my wildfire data to be a difficult source to analyze. One thing was researching to have a deeper insight into what exactly causes and affects wildfire. Another thing was analyzing and creating the correct visualization graphs and charts which can be used for my data exploration and insight collections. In my opinion, I think that although the research process had been extremely time-consuming and tedious, it was well worth it because the satisfaction of realizing and understanding how the data came to be was eye-opening.

Thirdly, I was not accustomed to using KNIME or Tableau. I had to relook at my previous labs, search for online videos and examples on how I could display and analyze my data while also using the applications and seek guidance and advice from my peers on how I could explain and show my data. After a long process of figuring out and using KNIME and Tableau, I am finally able to use these applications more easily and creatively.

Overall, I have learnt many things from this process of creating an analysis report. Now, I can have a better visualization of how graphs and charts work, as well as knowing how to use KNIME and Tableau. Understanding and realizing new things through experiences and obstacles are what will help me to grow as an individual and a future analyst.

**12. References**

[1] National Geographic, “Wildfires,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://education.nationalgeographic.org/resource/wildfires/>

[2] Ready Kids, “Wildfires,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://ready.gov/kids/disaster-facts/wildfires>

[3] World Health Organization, “Wildfires,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://who.int/health-topics/wildfires#tab=tab_1>

[4] EarthData, “Wildfires Data Pathfinder,” (n.d.) Accessed on: July 28, 2023 [Online].

Available: <https://www.earthdata.nasa.gov/learn/pathfinders/wildfires-data-pathfinder>

[5] Earth System Science Data, “A spatial database of wildfires in the United States, 1992-2011,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://essd.copernicus.org/articles/6/1/2014/>

[6] SAP, “Source System,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://help.sap.com/docs/SAP_NETWEAVER_750/ccc9cdbdc6cd4eceaf1e5485b1bf8f4b/801a618ae07211d2acb80000e829fbfe.html>

[7] NWCG, “NWCG,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.nwcg.gov/>

[8] NWCG, “NWCG Standards for Unit Identifiers PMS-931,” (n.d.) Accessed on: July 28, 2023 [Online].

Available: <https://www.nwcg.gov/sites/default/files/publications/pms931.pdf>

[9] NWCG, “Fire Size Class,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.nwcg.gov/data-standards/approved/fire-size-class>

[10] scientific reports, “Effects of ownership patterns on cross-boundary wildfires,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.nature.com/articles/s41598-021-98730-1>

[11] Bankrate, “2023 U.S. wildfire statistics,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.bankrate.com/insurance/homeowners-insurance/wildfire-statistics/#wildfire-statistics-by-state>

[12] Britannica, “wildfire,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.britannica.com/science/wildfire>

[13] little. America, “When is the best time to travel to the United States,” (n.d.) Accessed on: July 28, 2023 [Online].

Available: <https://www.littleamerica.co.uk/united-states/best-travel-time>

[14] Rainbow Restores, “When is Wildfire Season?,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://rainbowrestores.com/frequently-asked-questions/when-is-wildfire-season#:~:text=The%20peak%20month%20of%20wildfire,%2C%20Florida%2C%20Arizona%20and%20Oklahoma>.

[15] National Interagency Fire Center, “Fire Prevention, Education and Mitigation,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.nifc.gov/fire-information/fire-prevention-education-mitigation/wildfire-investigation#:~:text=Debris%20and%20open%20burning,-Burn%20barrels%20and&text=Debris%20burn%20escapes%20are%20one,specific%20causes%20of%20wildland%20fires>

[16] Los Angeles Times, “Is wildfire arson getting worse in California? Climate change doesn’t help,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.latimes.com/california/story/2021-12-31/is-wildfire-arson-getting-worse-in-california>

[17] CBS NEWS, “U.S. wildfires became larger, more frequent and more widespread in last 20 years,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.cbsnews.com/news/wildfires-us-study-larger-frequent-widespread-20-years/#:~:text=Wildfires%20in%20the%20U.S.%20have,from%20the%20past%20several%20decades>

[18] EOS, “U.S. Fires Quadrupled in Size, Tripled in Frequency in 20 Years,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://eos.org/articles/u-s-fires-quadrupled-in-size-tripled-in-frequency-in-20-years>

[19] College of Natural Resources News, “Explainer: How Wildfires Start and Spread,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://cnr.ncsu.edu/news/2021/12/explainer-how-wildfires-start-and-spread/>

[20] Lightning Eliminators, “Damage Caused by Lightning Strikes,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://www.lightningprotection.com/damage-caused-by-lightning-strikes/#:~:text=Fire%20damage%3A%20The%20biggest%20threat,causing%20complete%20damage%20to%20property>

[21] The DataFace, “When and Where are Wildfires Most Common in the U.S.?,” (n.d.) Accessed on: July 28, 2023 [Online]. Available: <https://thedataface.com/2018/11/public-health/wildfires-map>

[22] Forest Service U.S. Department of Agriculture, “Prescribed Fire,” (n.d.) Accessed on: July 28, 2023 [Online].

Available: <https://www.fs.usda.gov/managing-land/prescribed-fire>