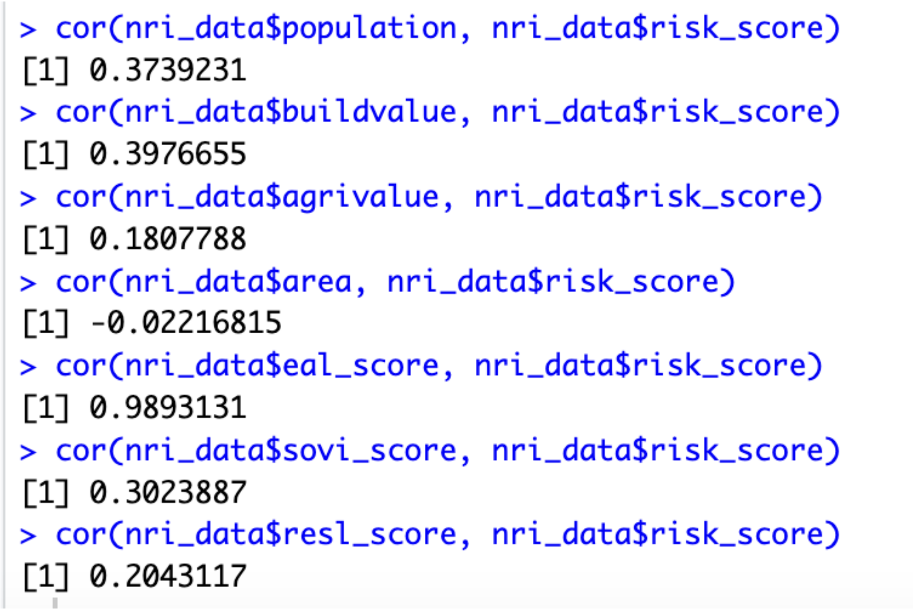
CORRELATION ANALYSIS

These correlation coefficients provide information about the strength and direction of the relationships between each pair of variables and the "risk_score." Here's a brief interpretation of the results:



```
> cor(nri_data$population, nri_data$risk_score)
[1] 0.3739231
> cor(nri_data$buildvalue, nri_data$risk_score)
[1] 0.3976655
> cor(nri_data$agrivalue, nri_data$risk_score)
[1] 0.1807788
> cor(nri_data$area, nri_data$risk_score)
[1] -0.02216815
> cor(nri_data$eal_score, nri_data$risk_score)
[1] 0.9893131
> cor(nri_data$sovi_score, nri_data$risk_score)
[1] 0.3023887
> cor(nri_data$resl_score, nri_data$risk_score)
[1] 0.2043117
```

Population, Building Value, and SOVI Score have positive correlations with Risk Score, indicating that higher values in these variables tend to be associated with higher risk scores.

Agriculture Value has a relatively weaker positive correlation with Risk Score.

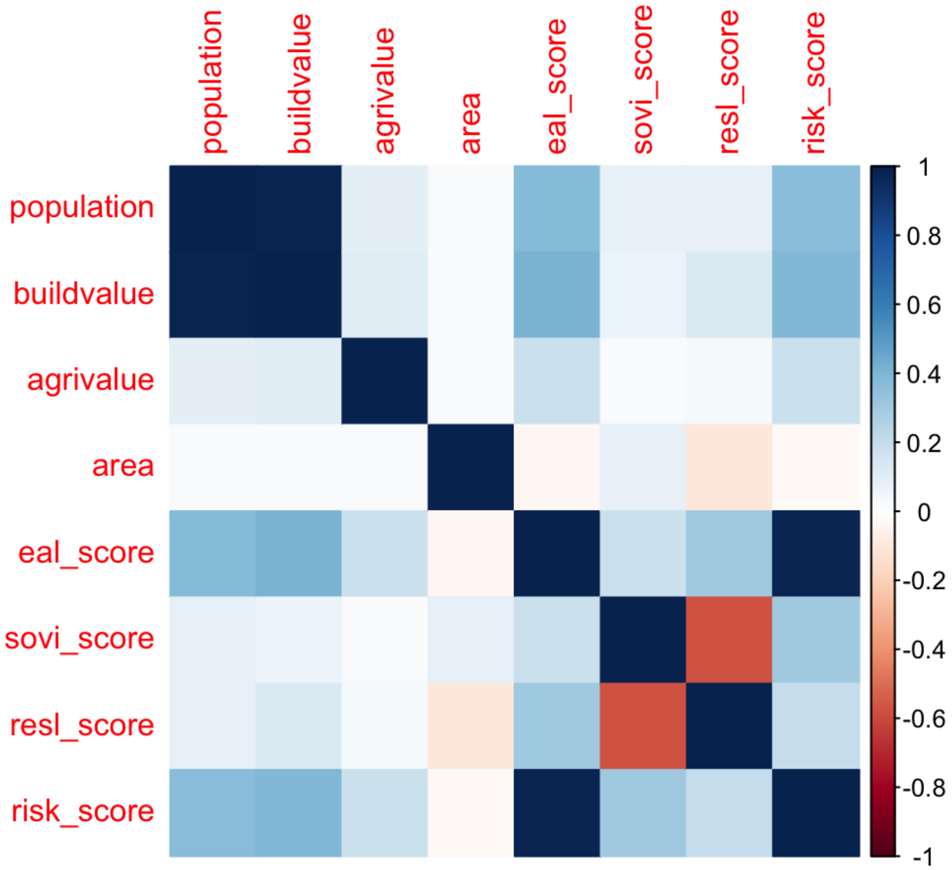
Area has a very weak negative correlation with Risk Score, suggesting that larger areas might be associated with slightly lower risk scores.

Expected Annual Loss (EAL) Score has a very strong positive correlation with Risk Score, indicating that as EAL Score increases, Risk Score increases significantly.

Social Vulnerability (SOVI) Score has a moderate positive correlation with Risk Score, suggesting that higher SOVI Scores are associated with higher risk scores.

Community Resilience (RESL) Score also has a positive correlation with Risk Score, but it's relatively weaker compared to other variables.

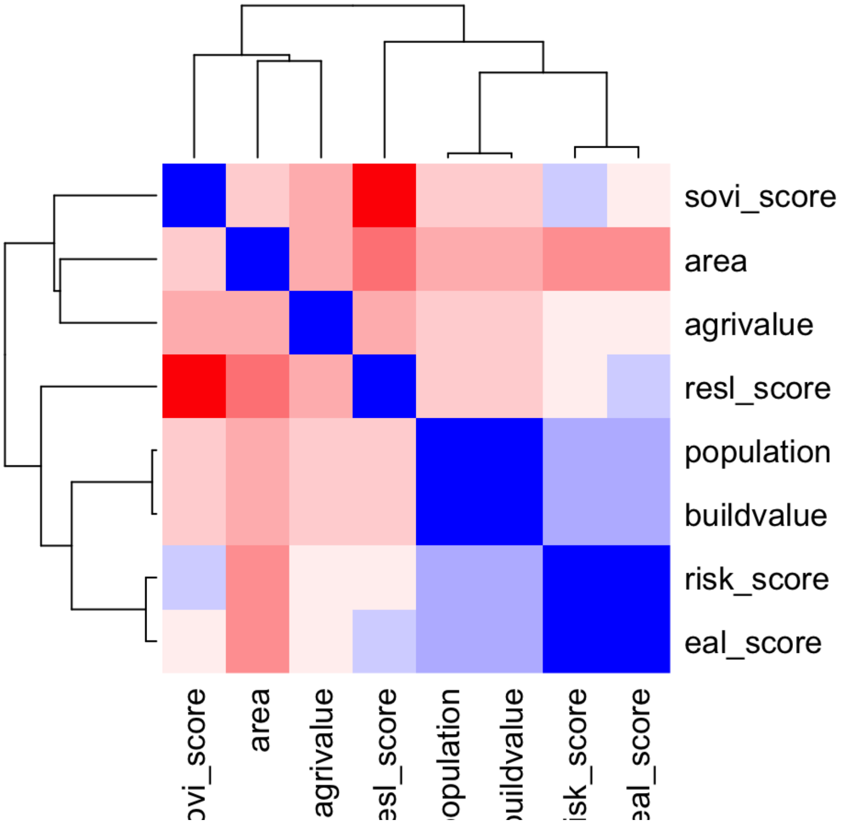
Overall, these correlations provide insights into how each variable is related to the overall risk score.



The intensity of the color in each cell of the plot represents the strength of the correlation between two variables. A darker color (e.g., dark blue) indicates a stronger correlation, while a lighter color (e.g., light yellow) indicates a weaker correlation.

The diagonal line from the top left to the bottom right typically contains perfect correlations (correlation of 1) because each variable is perfectly correlated with itself.

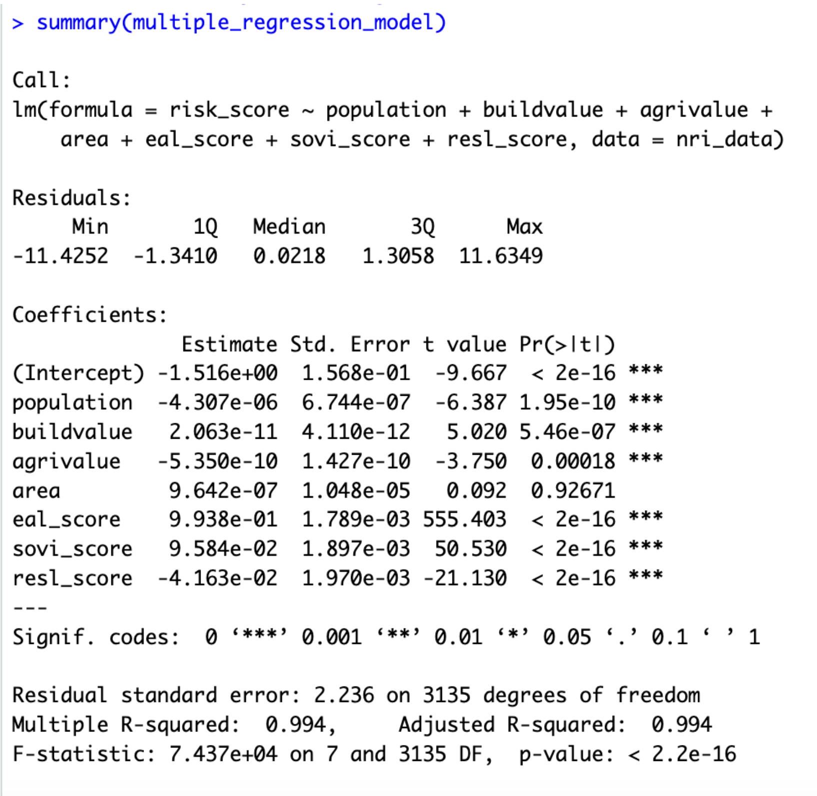
Variables with a positive correlation will have a positive color, typically in shades of blue. This means that as one variable increases, the other tends to increase as well. The darker the blue, the stronger the positive correlation. Variables with a negative correlation will have a negative color, typically in shades of red. This means that as one variable increases, the other tends to decrease. The darker the red, the stronger the negative correlation.



The heatmap created using the heatmap function and the corrplot visualization serve similar purposes in that they both display the correlations between variables in a dataset.

MULTIPLE LINEAR REGRESSION ANALYSIS

Since we are primarily interested in understanding the overall risk of natural disasters and its relationship with various factors (e.g., population, building value, agriculture value, etc.), we want to focus on the overall risk score as the dependent variable and include only relevant predictor variables that are likely to influence overall risk.

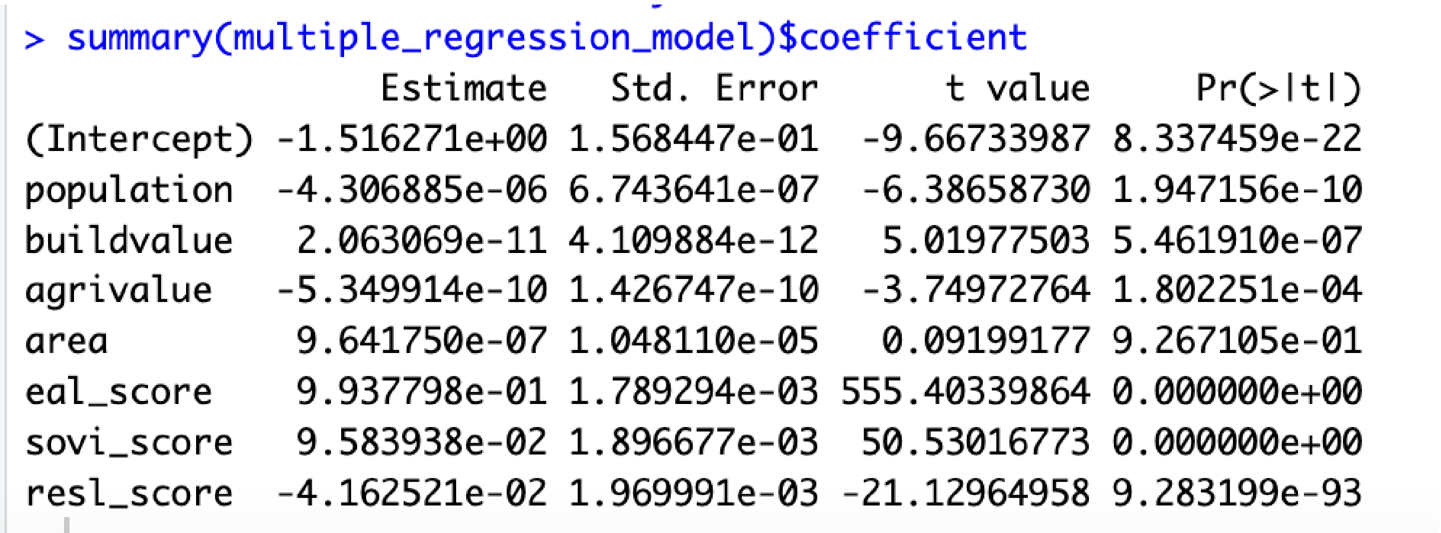


```
> summary(multiple_regression_model)

Call:
lm(formula = risk_score ~ population + buildvalue + agrivalue +
    area + eal_score + sovi_score + resl_score, data = nri_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.4252  -1.3410   0.0218   1.3058  11.6349 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.516e+00  1.568e-01  -9.667  < 2e-16 ***
population  -4.307e-06  6.744e-07  -6.387 1.95e-10 ***
buildvalue   2.063e-11  4.110e-12   5.020 5.46e-07 ***
agrivalue   -5.350e-10  1.427e-10  -3.750  0.00018 ***
area         9.642e-07  1.048e-05   0.092  0.92671    
eal_score    9.938e-01  1.789e-03 555.403  < 2e-16 ***
sovi_score   9.584e-02  1.897e-03  50.530  < 2e-16 ***
resl_score  -4.163e-02  1.970e-03 -21.130  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.236 on 3135 degrees of freedom
Multiple R-squared:  0.994,	Adjusted R-squared:  0.994 
F-statistic: 7.437e+04 on 7 and 3135 DF,  p-value: < 2.2e-16
```

The results from the multiple linear regression model provide insights into how various predictor variables are associated with the dependent variable, which is the "risk_score" in the analysis. Here's an interpretation of the key findings:

The "Pr(>|t|)" column provides p-values for each coefficient. In this summary, you can see that several predictor variables (population, buildvalue, agrivalue, eal_score, sovi_score, and resl_score) have p-values less than 0.05, indicating that they are statistically significant predictors of risk_score. The "Pr(>|t|)" column also shows that the intercept term is statistically significant.

The "Multiple R-squared" value is 0.994, indicating that the model explains a significant proportion of the variance in risk_score. This suggests that the combination of predictor variables in the model is highly effective in explaining the variation in the dependent variable.



```
> summary(multiple_regression_model)$coefficient
                 Estimate   Std. Error      t value     Pr(>|t|)
(Intercept) -1.516271e+00 1.568447e-01  -9.66733987 8.337459e-22
population  -4.306885e-06 6.743641e-07  -6.38658730 1.947156e-10
buildvalue   2.063069e-11 4.109884e-12   5.01977503 5.461910e-07
agrivalue   -5.349914e-10 1.426747e-10  -3.74972764 1.802251e-04
area         9.641750e-07 1.048110e-05   0.09199177 9.267105e-01
eal_score    9.937798e-01 1.789294e-03 555.40339864 0.000000e+00
sovi_score   9.583938e-02 1.896677e-03  50.53016773 0.000000e+00
resl_score  -4.162521e-02 1.969991e-03 -21.12964958 9.283199e-93
```

Population (Population Estimate): For every one-unit increase in population, the estimated "risk_score" decreases by approximately -4.31 units. This suggests that higher population areas tend to have lower risk scores, all else being equal.

Building Value (Buildvalue Estimate): For every one-unit increase in building value, the estimated "risk_score" increases by approximately 2.06 units. This implies that areas with higher building values tend to have higher risk scores, assuming other factors remain constant.

Agriculture Value (Agrivalue Estimate): For every one-unit increase in agriculture value, the estimated "risk_score" decreases by approximately -5.35 units. This suggests that areas with higher agricultural values tend to have lower risk scores, holding other variables constant.

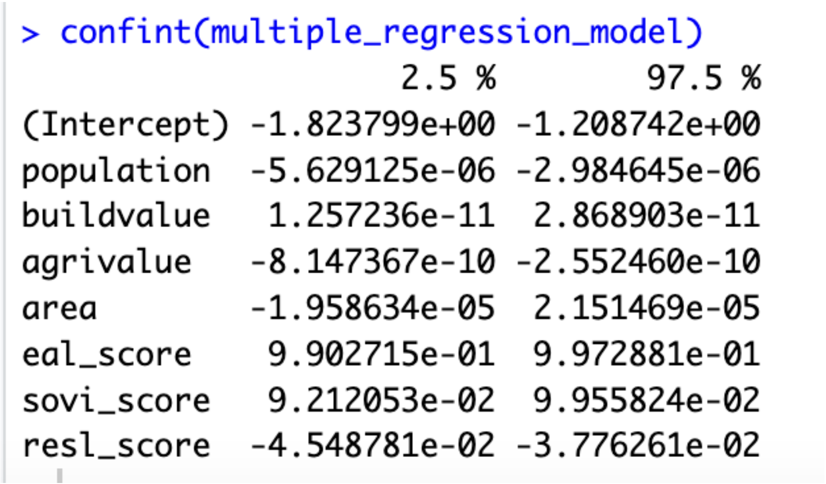
Area (Area Estimate): The "area" variable does not appear to have a significant effect on the "risk_score" because the coefficient estimate is close to zero (9.64e-07). In other words, changes in the area of an area do not appear to be strongly associated with changes in risk scores.

Expected Annual Loss Score (Eal_score Estimate): For every one-unit increase in the expected annual loss score (eal_score), the estimated "risk_score" increases by approximately 0.994 units. This suggests that higher EAL scores are associated with higher risk scores.

Social Vulnerability Score (Sovi_score Estimate): For every one-unit increase in the social vulnerability score (sovi_score), the estimated "risk_score" increases by approximately 0.096 units. This implies that higher social vulnerability scores are associated with higher risk scores.

Community Resilience Score (Resl_score Estimate): For every one-unit increase in the community resilience score (resl_score), the estimated "risk_score" decreases by approximately -0.042 units. This suggests that higher community resilience scores are associated with lower risk scores.

In summary, the multiple linear regression model provides estimates of how changes in each predictor variable are associated with changes in the "risk_score." Some variables, such as population, building value, agriculture value, social vulnerability score, and community reliance score, appear to have statistically significant associations with the "risk_score," while others, like area, do not appear to be significant in this model.



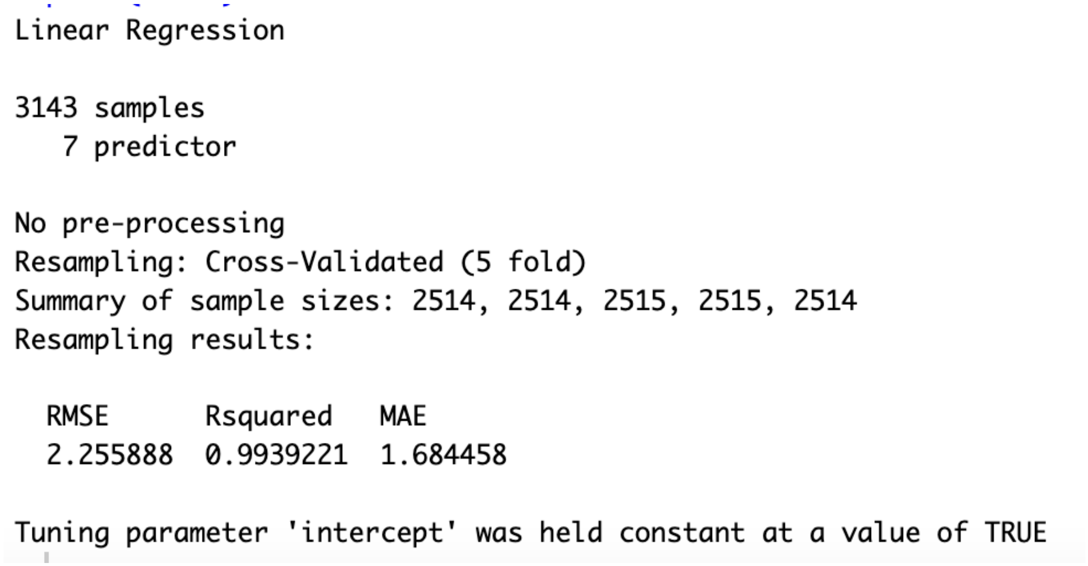
```
> confint(multiple_regression_model)
                    2.5 %        97.5 %
(Intercept) -1.823799e+00 -1.208742e+00
population  -5.629125e-06 -2.984645e-06
buildvalue   1.257236e-11  2.868903e-11
agrivalue   -8.147367e-10 -2.552460e-10
area        -1.958634e-05  2.151469e-05
eal_score    9.902715e-01  9.972881e-01
sovi_score   9.212053e-02  9.955824e-02
resl_score  -4.548781e-02 -3.776261e-02
```

area: The confidence interval for the area coefficient includes zero. This suggests that the area variable may not have a statistically significant effect on the risk_score. In other words, changes in the area may not be associated with changes in the risk_score in a statistically meaningful way, at least based on this model.

All other variables have confidence intervals for these coefficients that do not include zero, indicating that these variables have statistically significant effects on the risk_score.

VALIDATION ANALYSIS

The cross-validated results for the multiple linear regression model are as follows:



```
Linear Regression

3143 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 2514, 2514, 2515, 2515, 2514
Resampling results:

  RMSE      Rsquared   MAE
  2.255888  0.9939221  1.684458

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Root Mean Squared Error (RMSE): Approximately 2.255888
R-squared (Rsquared): Approximately 0.9939221
Mean Absolute Error (MAE): Approximately 1.684458
These performance metrics provide insights into how well the regression model is performing:

RMSE: This metric measures the average prediction error of the linear regression model. An RMSE of approximately 2.255888 means that, on average, the predicted risk_score values from the model are off by around 2.255888 units from the actual risk_score values in the dataset.

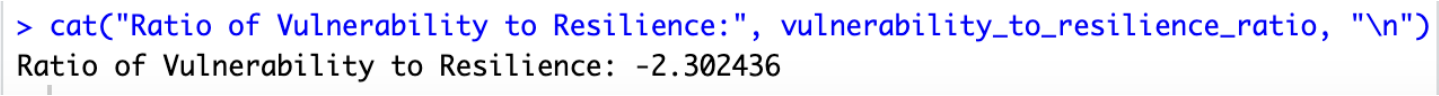
R-squared (Rsquared): R-squared measures the goodness of fit of the linear regression model. An R-squared value of approximately 0.9939221 indicates that the model can explain approximately 99.39% of the variance in the risk_score variable. In other words, the model does an excellent job of capturing and explaining the variability in risk_score based on the predictor variables.

MAE: MAE represents the average absolute difference between the model's predictions and the actual risk_score values. An MAE of approximately 1.684458 means that, on average, the model's predictions are off by about 1.684458 units from the actual values.

In summary, the RMSE and MAE values are relatively low, indicating that the regression model is making accurate predictions. The R-squared value is very high, suggesting that the model explains a significant portion of the variance in the risk_score, indicating strong predictive power. These results suggest that the multiple linear regression model is performing very well in predicting the risk_score based on the chosen predictor variables. It has a high degree of explanatory power (as indicated by the high R-squared value) and makes relatively accurate predictions (as indicated by the low RMSE and MAE values).

VULNERABILITY / RESILIENCE RATIO

Calculating the vulnerability-to-resilience ratio can provide valuable insights into disaster risk assessment and preparedness. This ratio helps assess the balance between a community's vulnerability (its susceptibility to negative impacts from disasters) and its resilience (its ability to bounce back and recover from disasters).



```
> cat("Ratio of Vulnerability to Resilience:", vulnerability_to_resilience_ratio, "\n")
Ratio of Vulnerability to Resilience: -2.302436
```

I used the coefficients from the multiple regression model to calculate the ratio of vulnerability to resilience because these coefficients represent the estimated effect of each predictor variable on the outcome variable (risk_score) while holding other variables constant.

In this context, the coefficients tell us how much the risk score is expected to change for a one-unit change in the predictor variable, assuming all other variables remain constant. By dividing the coefficient of sovi_score by the coefficient of resl_score, we are essentially calculating the change in risk score associated with a one-unit change in social vulnerability relative to a one-unit change in community resilience.

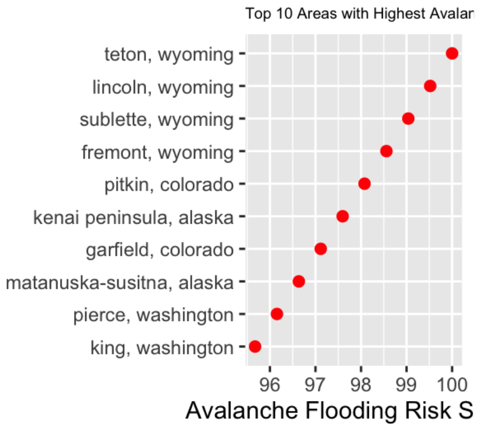
So, when I calculated the ratio of vulnerability to resilience using the coefficients, I was quantifying the relative impact of changes in social vulnerability (sovi_score) compared to changes in community resilience (resl_score) on the risk score. This ratio provides insights into the relationship between vulnerability and resilience in the context of the regression model.

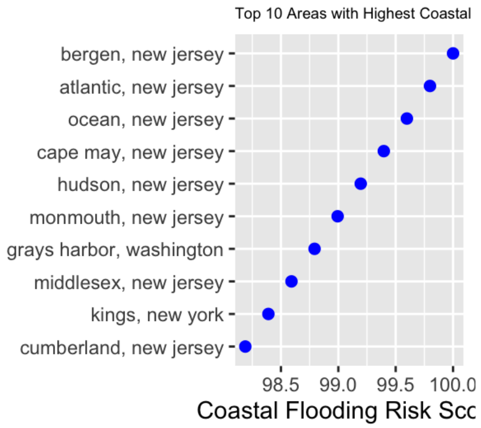
The calculated ratio of vulnerability to resilience is approximately -2.302436. This ratio suggests that for every one-unit increase in the social vulnerability score (sovi_score), there is an expected decrease of approximately 2.302 units in the resilience score (resl_score) based on the multiple regression model.

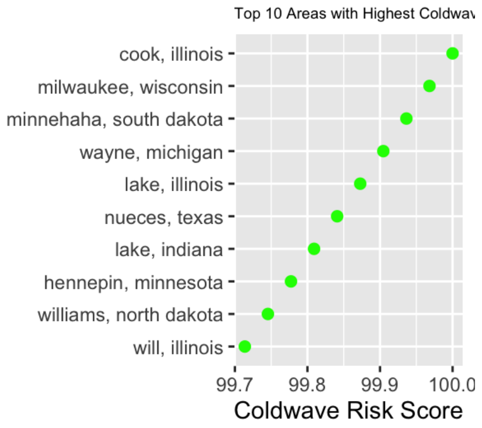
A negative ratio like this implies that as vulnerability increases, resilience tends to decrease, which is an important insight for assessing the impact of social vulnerability on community resilience in the context of natural disaster risk.

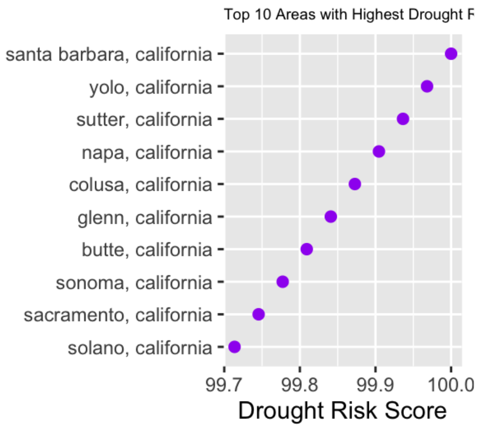
FACETTED SCATTERPLOTS OF THE TOP 10 COUNTIES PER NATURAL DISASTER SCORE

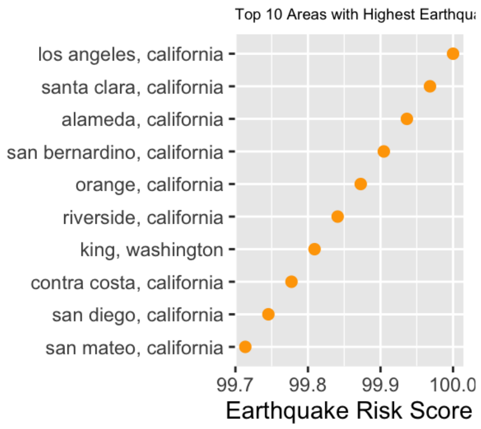
The code creates individual scatter plots for different types of natural disasters (e.g., Avalanche, Coastal Flooding, Coldwave, etc.). Each plot uses the ggplot function to create a scatter plot (geom_point) where the x-axis represents the risk score for the specific natural disaster, and the y-axis represents the county names (reordered based on risk score). The color of the points in each plot corresponds to the specific natural disaster (e.g., red for Avalanche, blue for Coastal Flooding, etc.). The title of each plot indicates the disaster type and mentions that it represents the top 10 areas with the highest risk scores for that particular disaster.
These scatter plots help visually assess the distribution and variation in risk scores across different counties for each natural disaster. Comparisons can be made between different disasters and their impact on various counties. Overall, this set of scatter plots is a powerful tool for quickly grasping the relative risk scores across different counties and disasters. It facilitates the identification of counties with the highest risk scores for each disaster type.

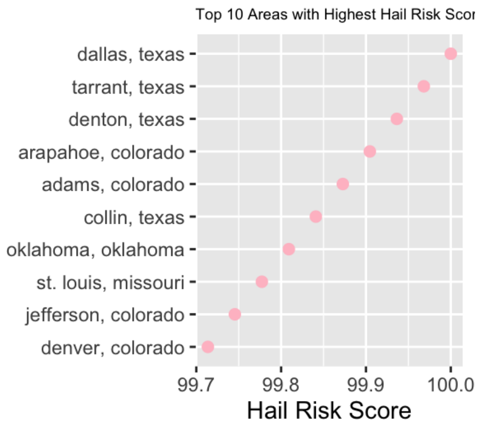

Top 10 Areas with Highest Avalar

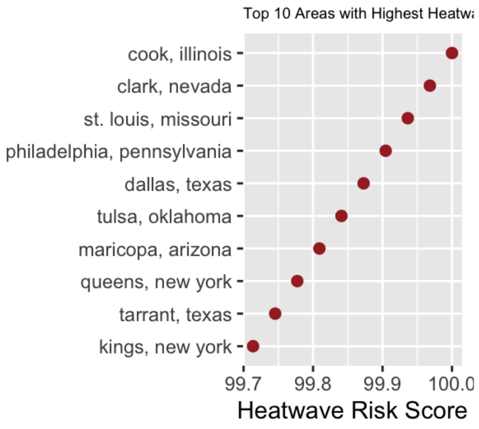

Top 10 Areas with Highest Coastal

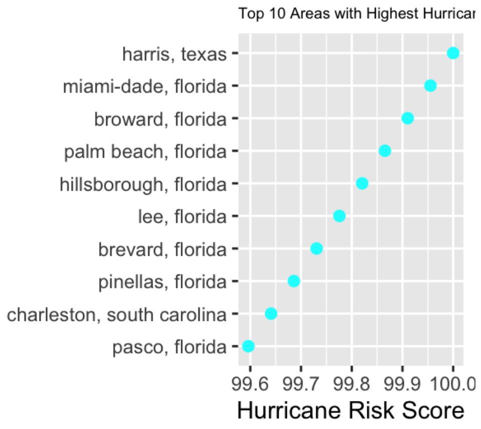

Top 10 Areas with Highest Coldwav

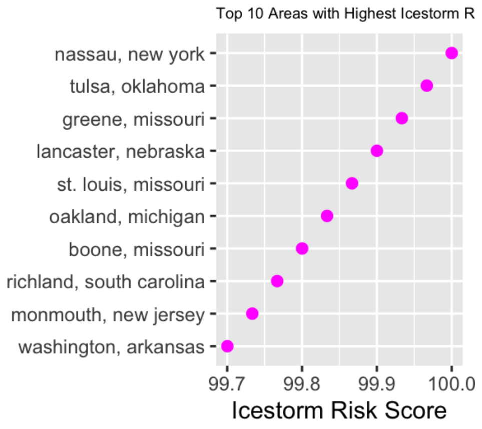

Top 10 Areas with Highest Drought F

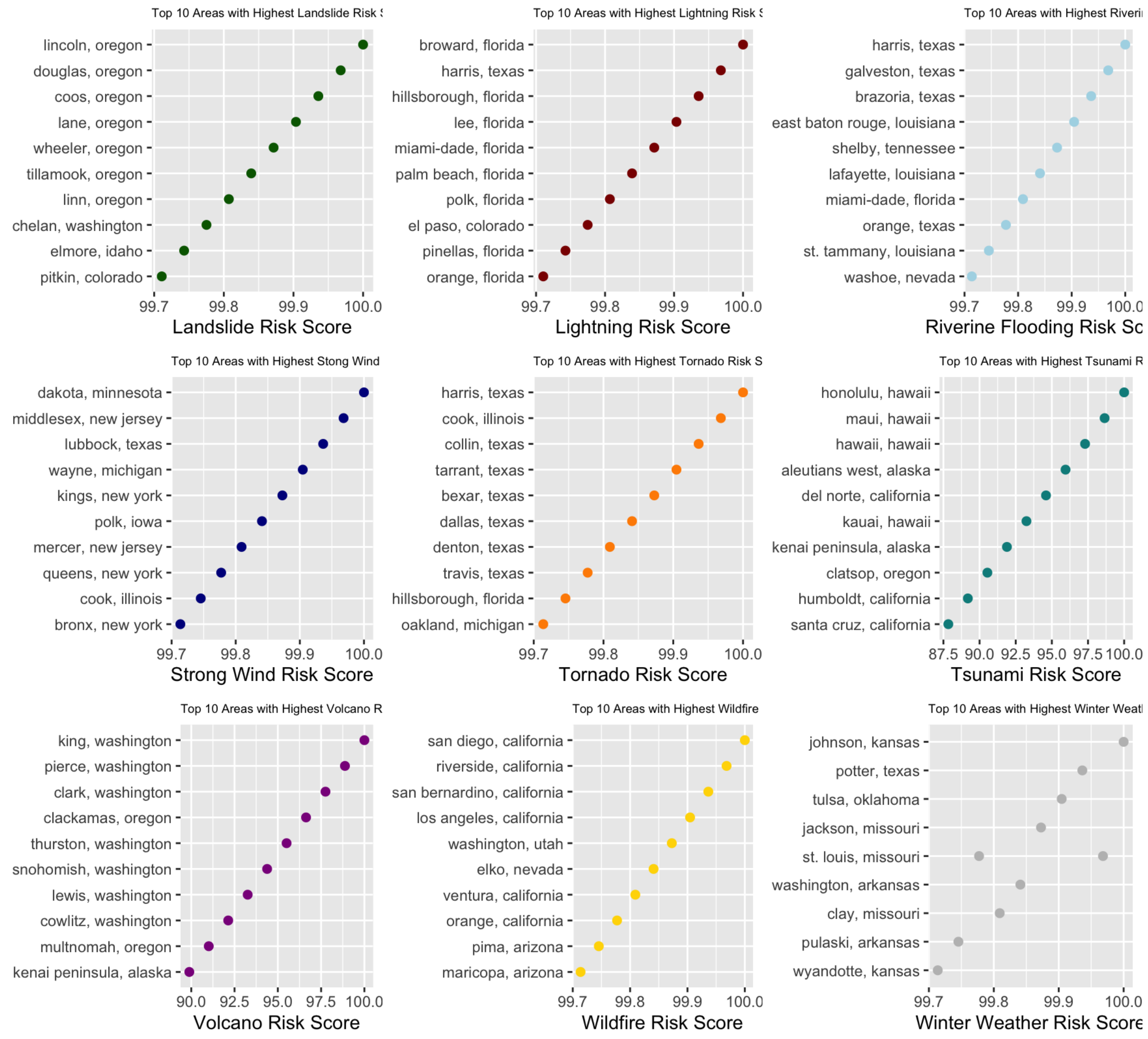

Top 10 Areas with Highest Earthqu


Top 10 Areas with Highest Hail Risk Scor


Top 10 Areas with Highest Heatwa

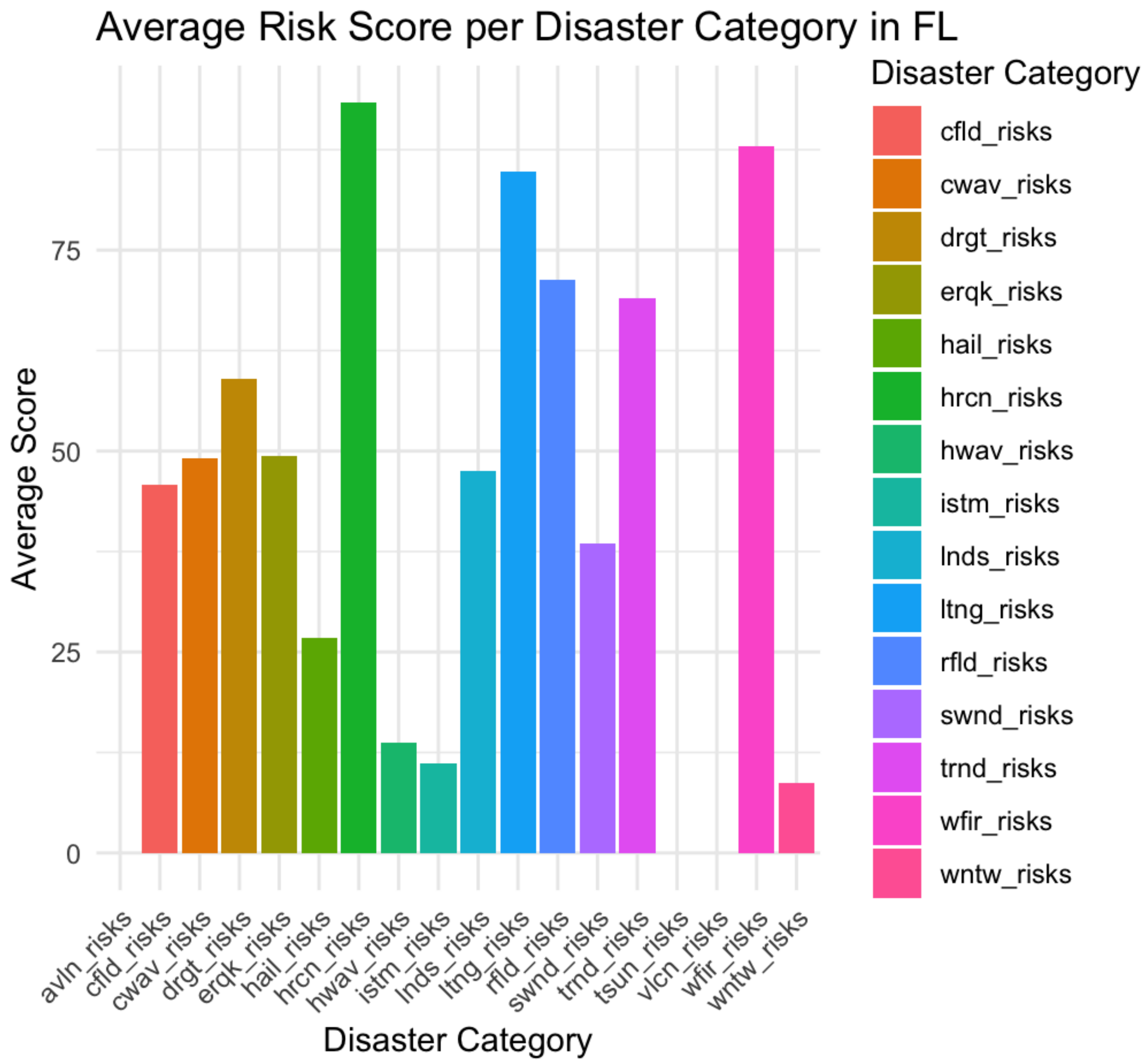

Top 10 Areas with Highest Hurricar


Top 10 Areas with Highest Icestorm R



BAR CHARTS OF DISASTER RISK

The bar chart created visualizes the average risk scores for different disaster categories in each state of the US. For each state, you can identify which disaster categories contribute the most to the overall risk profile. This insight can guide prioritization in terms of preparedness and mitigation efforts. You can also compare the average risk scores across different states to identify states with higher or lower overall risk levels. This can be valuable for regional risk assessments and resource allocation.

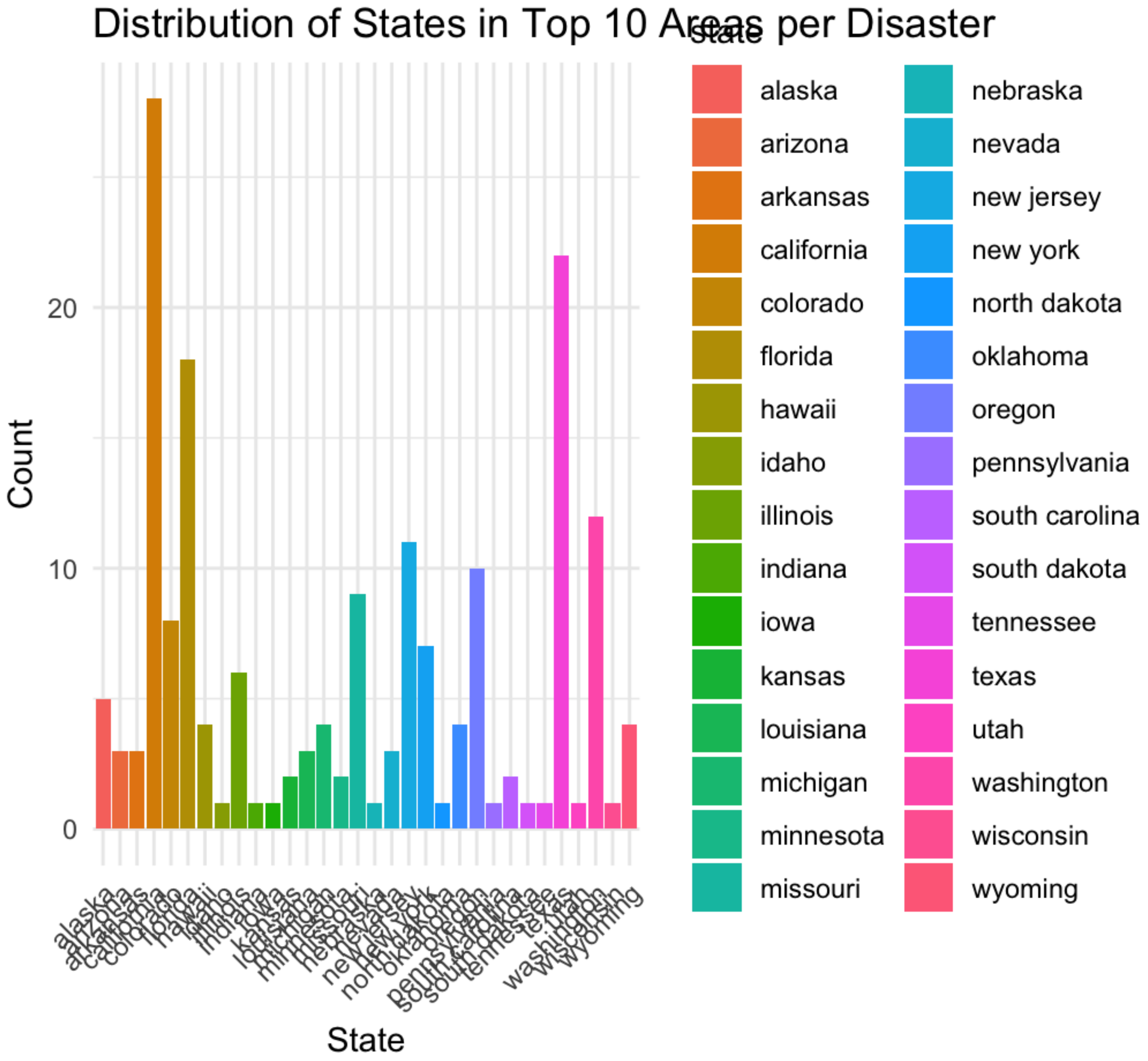


BAR CHARTS OF DISTRIBUTIONS OF TOP 10 AREAS PER DISASTER

Here, I created a bar chart to visualize the distribution of states in the top 10 areas per disaster. The height of each bar represents the count (frequency) of states that appear in the top 10 areas for different types of disasters. This gives an indication of how often each state is mentioned as a high-risk area across various disaster categories. Different colors represent different types of disasters (e.g., hurricanes, wildfires, earthquakes). By examining the distribution of states across these colors, you can observe which states are frequently mentioned as high-risk areas for specific types of disasters.

California as a High-Risk State: The fact that California ranks highest suggests that it frequently appears in the top 10 areas for a variety of disasters. This aligns with California's diverse geography and susceptibility to various natural hazards, including wildfires, earthquakes, and coastal events.
Texas and Florida's Presence: Texas and Florida being the second and third highest, respectively, indicates that these states are also prominently featured in the top areas for various types of disasters. Texas, for

example, is prone to hurricanes, flooding, and severe weather events. Florida faces risks from hurricanes, flooding, and coastal hazards.



BUBBLE MAP

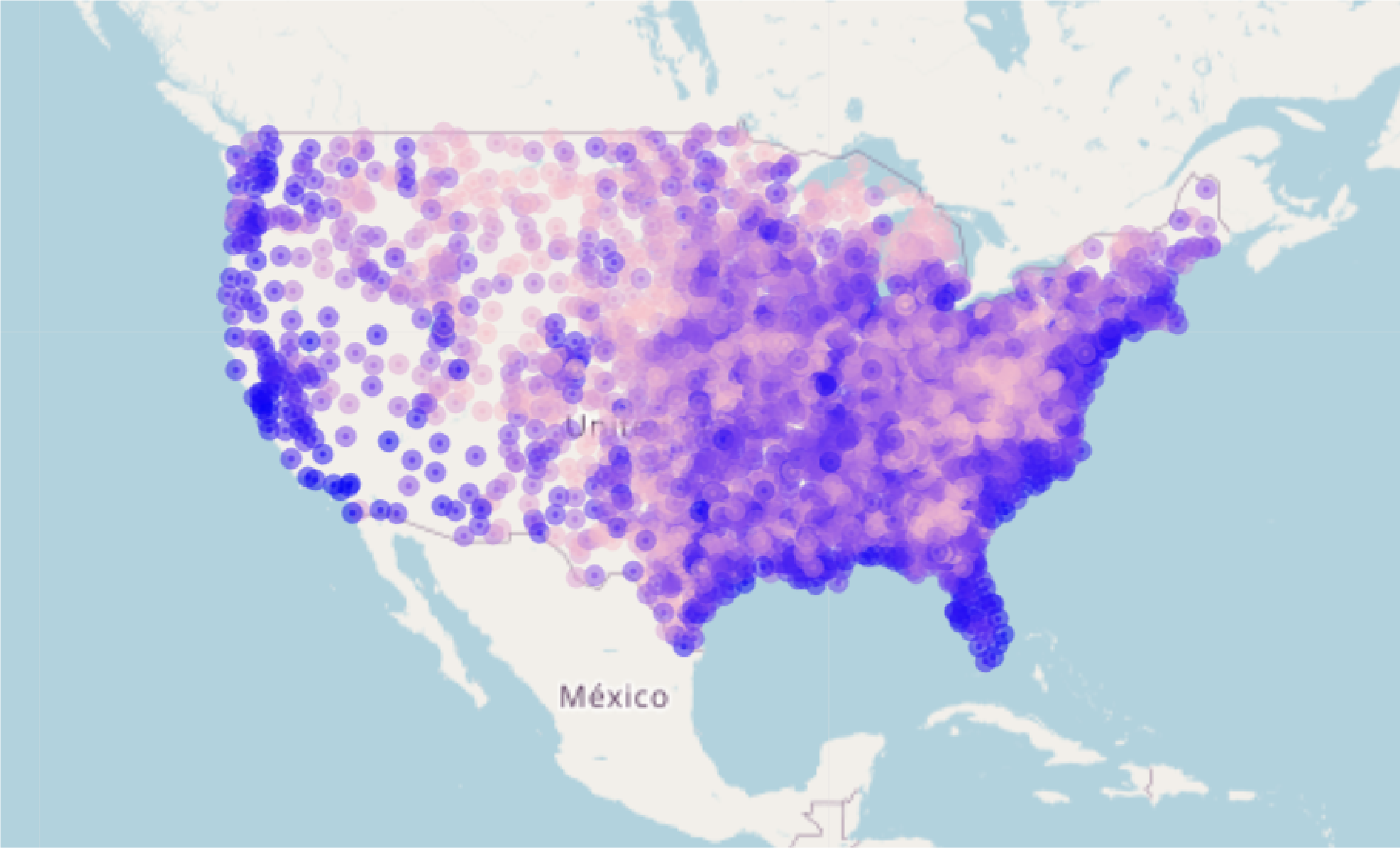
The provided R code utilizes the leaflet package to create interactive maps that visualize the geographic distribution of risk scores across counties and states. Here's what each section of the code shows:

The first map focuses on Alabama, showing circles for each county with the risk_score as the size of the circles.

The second map covers the entire United States, iterating over each state and adding circles with risk_score information.

The third map introduces a color gradient (from pink to blue) based on risk_score values.

These visualizations help in understanding the geographical distribution and risk variation across counties and states. The color-coded map, in particular, allows for a quick assessment of risk levels based on the risk_score.



CHOROPLETH MAP

The code created generates a choropleth map, a type of thematic map in which areas are shaded or patterned in proportion to the value of a variable being represented. In this case, the choropleth map represents the spatial distribution of risk scores across different counties.

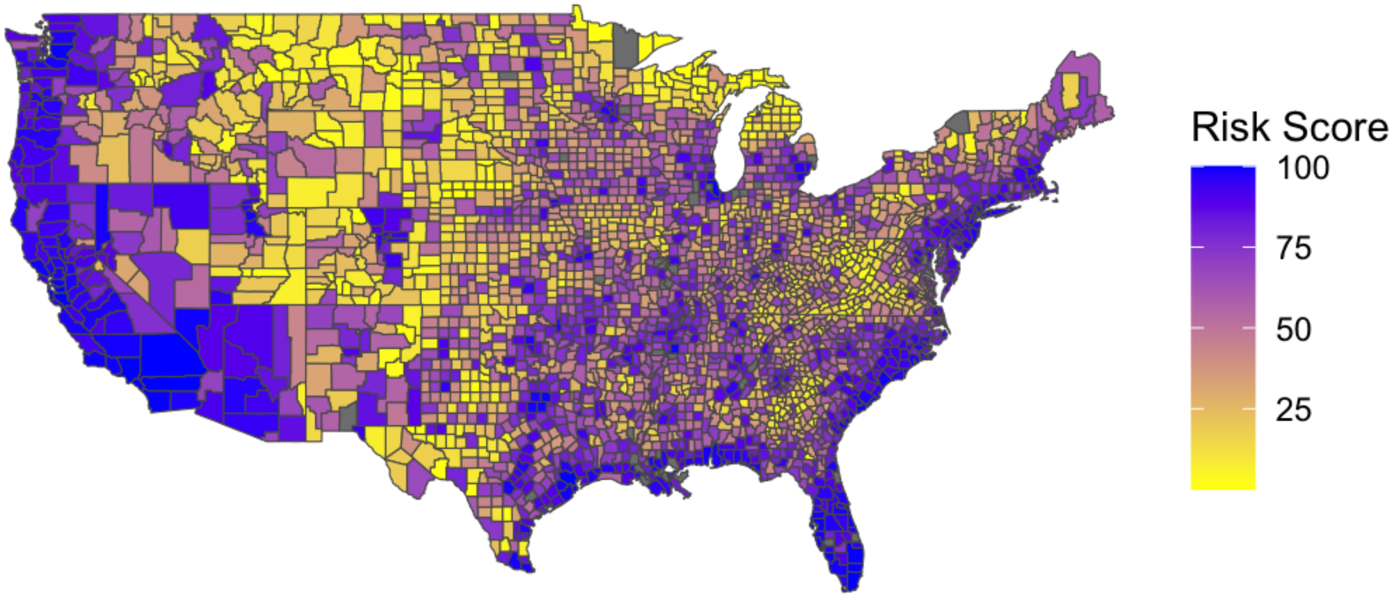
This choropleth map visually represents the spatial distribution of risk scores across counties.
Counties are colored based on their respective risk scores, allowing for quick identification of areas with higher or lower risk levels.

Geographic Distribution of Risk Scores: The map visually communicates the geographic distribution of risk scores across different counties. Darker shades of blue may indicate higher risk scores, while lighter shades represent lower risk scores.

Regional Patterns: Patterns or clusters of similar risk scores may be identified across regions, helping to understand the spatial variation in risk levels.

Identification of High and Low-Risk Areas: The choropleth map allows for the quick identification of areas with higher or lower risk scores, providing valuable insights for risk assessment and decision-making.

Geospatial Context: By visualizing risk scores on a map, stakeholders can gain a better understanding of the spatial context of risk, facilitating targeted interventions and resource allocation.
In summary, the choropleth map provides a spatial overview of risk scores, allowing for a more intuitive understanding of the distribution of risk across different geographic areas.



HEXABIN MAP

The provided code creates visualizations that show the spatial distribution of risk scores across U.S. states using hexagonal bins. The first set of maps provides a visual representation of the hexgrid with state names and abbreviations. The final chloropleth hexabin map overlays the NRI data (specifically risk_score) on the hexgrid, offering insights into the spatial distribution of risk scores across states.

Insights:

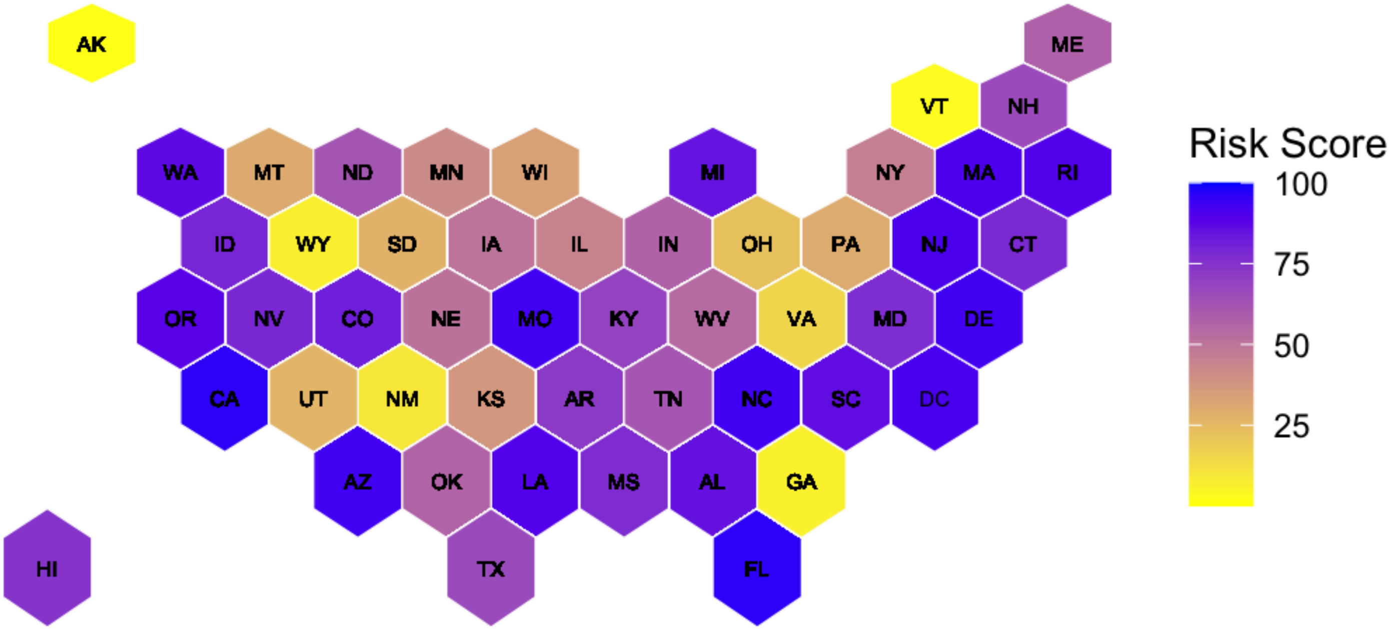
Spatial Distribution of Risk Scores: The choropleth hexabin map provides a visual representation of how risk scores are distributed across U.S. states. Darker hexagons represent areas with higher risk scores, while lighter hexagons represent areas with lower risk scores.

State Labels: Hexagon centroids have state abbreviations labeled on them, providing context to the viewer about which state each hexagon represents.

Comparison of Risk Scores: The map allows for a quick comparison of risk scores between different states. Patterns and clusters of higher or lower risk areas may be identifiable.

Geospatial Context: The hexagon grid format helps maintain a consistent spatial representation, making it easier to compare risk scores across states.

In summary, the visualizations provide insights into the spatial distribution of risk scores, allowing viewers to identify areas with varying levels of risk across U.S. states. The use of hexagonal bins facilitates a more structured and visually appealing representation of the data.


AK
ME
VT
NH
WA
MT
ND
MN
WI
MI
NY
MA
RI
ID
WY
SD
IA
IL
IN
OH
PA
NJ
CT
OR
NV
CO
NE
MO
KY
WV
VA
MD
DE
CA
UT
NM
KS
AR
TN
NC
SC
DC
AZ
OK
LA
MS
AL
GA
HI
TX
FL
Risk Score
100
75
50
25