```r
fire_data<-read.csv("https://raw.githubusercontent.com/aarongelf/data602-data/refs/heads/main/fp-histor
#We have also include the dataset in our submission, in case there is an error accessing the URL.
```

Question 1 - What is the Relationship Between Wind Speed and Fire Spread Rate

To examine the relationship between wind speed and fire spread rate, we can perform a few visual functions, to get an idea of the dataset itself. It is also important to note that data dictionary provided should be used to help interpret the columns based on their names, as well as what the values represent. The data dictionary can be found at https://open.alberta.ca/dataset/a221e7a0-4f46-4be7-9c5a-e29de9a3447e/resource/1b635b8b-a937-4be4-857e-8aeef77365d2/download/fp-historical-wildfire-data-dictionary-2006-2023.pdf.

```r
head(fire_data)
```

```
##   fire_year fire_number fire_name current_size size_class
## 1      2006      PWF001      <NA>         0.10          A
## 2      2006      EWF002      <NA>         0.20          B
## 3      2006      EWF001      <NA>         0.50          B
## 4      2006      EWF003      <NA>         0.01          A
## 5      2006      PWF002      <NA>         0.10          A
## 6      2006      CWF001      <NA>         0.20          B
##   fire_location_latitude fire_location_longitude        fire_origin
## 1               56.24996               -117.1820       Private Land
## 2               53.60637               -115.9157    Provincial Land
## 3               53.61093               -115.5943    Provincial Land
## 4               53.60887               -115.6095    Provincial Land
## 5               56.24996               -117.0502    Provincial Land
## 6               51.15293               -115.0346 Indian Reservation
##   general_cause_desc industry_identifier_desc        responsible_group_desc
## 1           Resident                     <NA>                      Resident
## 2         Incendiary                     <NA> Others (explain in remarks)
## 3         Incendiary                     <NA> Others (explain in remarks)
## 4         Incendiary                     <NA> Others (explain in remarks)
## 5     Other Industry          Waste Disposal                     Employees
## 6           Resident                     <NA>                      Resident
##   activity_class        true_cause   fire_start_date det_agent_type det_agent
## 1          Grass    Permit Related  2006-04-02 12:00             UNP       310
## 2 Lighting Fires  Arson Suspected  2006-04-03 12:10             UNP       310
## 3 Lighting Fires  Arson Suspected  2006-04-03 12:15             UNP       310
## 4 Lighting Fires  Arson Suspected  2006-04-03 12:10             UNP       PUB
## 5         Refuse    Permit Related  2006-04-03 17:00             UNP       LFS
## 6   Unclassified       Unsafe Fire 2006-04-02 14:25             UNP       310
##     discovered_date discovered_size    reported_date dispatched_resource
## 1              <NA>              NA 2006-04-02 20:46           FPD Staff
## 2              <NA>              NA 2006-04-03 12:27           FPD Staff
## 3              <NA>              NA 2006-04-03 12:36           FPD Staff
## 4              <NA>              NA 2006-04-03 13:23           FPD Staff
## 5 2006-04-03 19:11              NA 2006-04-03 19:12           FPD Staff
## 6 2006-04-02 14:27              NA 2006-04-02 14:30           FPD Staff
##      dispatch_date start_for_fire_date assessment_resource assessment_datetime
## 1 2006-04-02 21:10    2006-04-02 21:20           IA Forces    2006-04-02 22:00
## 2 2006-04-03 12:33    2006-04-03 12:35           IA Forces    2006-04-03 13:20
## 3 2006-04-03 12:36    2006-04-03 12:42           IA Forces    2006-04-03 13:23
```

```
## 4 2006-04-03 13:50      2006-04-03 13:50            IA Forces      2006-04-03 14:08
## 5 2006-04-03 19:19      2006-04-03 19:22                Other      2006-04-03 19:57
## 6 2006-04-02 14:40      2006-04-02 14:45            IA Forces      2006-04-02 16:00
##   assessment_hectares  fire_spread_rate fire_type  fire_position_on_slope
## 1                0.01               0.0   Surface                    Flat
## 2                0.20               0.0   Surface               Lower 1/3
## 3                0.50               0.0   Surface                  Bottom
## 4                0.01               0.0   Surface                    Flat
## 5                0.10               0.1   Surface                    Flat
## 6                0.20               0.0   Surface                    Flat
##   weather_conditions_over_fire  temperature  relative_humidity  wind_direction
## 1                        Clear           18                 10              SW
## 2                        Clear           12                 22              SW
## 3                        Clear           12                 22              SW
## 4                        Clear           12                 22              SW
## 5                        Clear            6                 37              SW
## 6                        Clear           11                 32               S
##   wind_speed  fuel_type initial_action_by  ia_arrival_at_fire_date  ia_access
## 1          2        O1a        Land Owner                     <NA>       <NA>
## 2         10        O1a   Fire Department                     <NA>       <NA>
## 3         10        O1a   Fire Department                     <NA>       <NA>
## 4         10        O1b          Industry                     <NA>       <NA>
## 5          2       <NA>   Fire Department                     <NA>       <NA>
## 6         20        O1b   Fire Department                     <NA>       <NA>
##   fire_fighting_start_date  fire_fighting_start_size  bucketing_on_fire
## 1                     <NA>                        NA               <NA>
## 2                     <NA>                        NA               <NA>
## 3                     <NA>                        NA               <NA>
## 4                     <NA>                        NA               <NA>
## 5                     <NA>                        NA               <NA>
## 6                     <NA>                        NA               <NA>
##   distance_from_water_source  first_bucket_drop_date          bh_fs_date
## 1                         NA                    <NA>  2006-04-02 22:00
## 2                         NA                    <NA>  2006-04-03 13:20
## 3                         NA                    <NA>  2006-04-03 13:23
## 4                         NA                    <NA>  2006-04-03 14:08
## 5                         NA                    <NA>  2006-04-03 19:57
## 6                         NA                    <NA>  2006-04-02 16:00
##   bh_hectares          uc_fs_date  uc_hectares          to_fs_date  to_hectares
## 1        0.01   2006-04-02 22:00         0.01                <NA>           NA
## 2        0.20   2006-04-03 13:20         0.20                <NA>           NA
## 3        0.50   2006-04-03 13:23         0.50                <NA>           NA
## 4        0.01   2006-04-03 14:08         0.01                <NA>           NA
## 5        0.10  2006-04-03 20:19          0.10 2006-04-03 20:20          0.1
## 6        0.20   2006-04-02 16:00         0.20                <NA>           NA
##           ex_fs_date  ex_hectares
## 1 2006-04-03 10:20          0.10
## 2 2006-04-03 14:00          0.20
## 3 2006-04-03 15:00          0.50
## 4 2006-04-03 15:05          0.01
## 5 2006-04-05 10:18          0.10
## 6 2006-04-03 18:00          0.20
```

```r
colnames(fire_data)
```

```
##  [1] "fire_year"                    "fire_number"
##  [3] "fire_name"                    "current_size"
##  [5] "size_class"                   "fire_location_latitude"
##  [7] "fire_location_longitude"      "fire_origin"
##  [9] "general_cause_desc"           "industry_identifier_desc"
## [11] "responsible_group_desc"       "activity_class"
## [13] "true_cause"                   "fire_start_date"
## [15] "det_agent_type"               "det_agent"
## [17] "discovered_date"              "discovered_size"
## [19] "reported_date"                "dispatched_resource"
## [21] "dispatch_date"                "start_for_fire_date"
## [23] "assessment_resource"          "assessment_datetime"
## [25] "assessment_hectares"          "fire_spread_rate"
## [27] "fire_type"                    "fire_position_on_slope"
## [29] "weather_conditions_over_fire" "temperature"
## [31] "relative_humidity"            "wind_direction"
## [33] "wind_speed"                   "fuel_type"
## [35] "initial_action_by"            "ia_arrival_at_fire_date"
## [37] "ia_access"                    "fire_fighting_start_date"
## [39] "fire_fighting_start_size"     "bucketing_on_fire"
## [41] "distance_from_water_source"   "first_bucket_drop_date"
## [43] "bh_fs_date"                   "bh_hectares"
## [45] "uc_fs_date"                   "uc_hectares"
## [47] "to_fs_date"                   "to_hectares"
## [49] "ex_fs_date"                   "ex_hectares"
```

Based on initial inspection, we are only interested in a few columns, therefore we will create a new data frame, focusing on variables that will help examine the relationship between wind speed and fire spread rate.

```r
fire_data_1=fire_data %>%
  select(wind_speed,fire_spread_rate)
sum(is.na(fire_data_1))
```

```
## [1] 5575
```

```r
sum(sapply(fire_data_1[c("fire_spread_rate","wind_speed")],is.na))
```

```
## [1] 5575
```

```r
#Remove rows with NA values.   We won't include fuel_type for this part, as we are not initially concern
fire_clean=na.omit(fire_data_1[,c("wind_speed","fire_spread_rate")])
sum(is.na(fire_clean))
```

```
## [1] 0
```

```r
summary(fire_clean)
```

```
##    wind_speed      fire_spread_rate
##  Min.    : 0.000   Min.    : -1.0000
##  1st Qu.: 3.000    1st Qu.:  0.0000
##  Median : 6.000    Median :  0.0000
##  Mean   : 8.813    Mean   :  0.8962
##  3rd Qu.:12.000    3rd Qu.:  1.0000
##  Max.   :90.000    Max.   :100.0000
```

Based on our summary statistics, we can see that the minimum value for fire_spread_rate is -1. This seems peculiar, and warrants a bit of further investigation, therefore we will go back to the original data set and inspect any rows where fire_spread_rate is -1.

```r
negative_fire_spread_data <- fire_data[!is.na(fire_data$fire_spread_rate) & fire_data$fire_spread_rate
head(negative_fire_spread_data)
```

```
##       fire_year fire_number fire_name current_size size_class
## 12962      2014      RWF022      <NA>         0.20          B
## 13717      2015      HWF100      <NA>         0.02          A
## 18213      2017      MWF091      <NA>         0.10          A
## 20474      2020      CWF038      <NA>         0.01          A
## 20778      2019      SWF063      <NA>         0.01          A
## 21355      2019      SWF092      <NA>         0.04          A
##       fire_location_latitude fire_location_longitude        fire_origin
## 12962               52.37868               -115.6254     Provincial Land
## 13717               58.48312               -114.4696 Indian Reservation
## 18213               56.83563               -111.7322     Provincial Land
## 20474               51.10135               -115.3091     Provincial Land
## 20778               55.94107               -113.7807 Indian Reservation
## 21355               56.79307               -114.7125     Provincial Land
##       general_cause_desc industry_identifier_desc          responsible_group_desc
## 12962       Undetermined                     <NA>                            <NA>
## 13717         Incendiary                     <NA>                            <NA>
## 18213         Recreation                     <NA> Others (explain in remarks)
## 20474         Recreation                     <NA>                         Campers
## 20778         Incendiary                     <NA>                            <NA>
## 21355          Lightning                     <NA>                            <NA>
##             activity_class          true_cause   fire_start_date det_agent_type
## 12962                 <NA>                <NA>  2014-05-24  4:00            LKT
## 13717         Unclassified                <NA> 2015-05-18 10:48            LKT
## 18213        OHV Operation Burning Substance 2017-09-16 16:38            AIR
## 20474 Cooking and Warming         Unsafe Fire 2020-06-27 18:00            UNP
## 20778                Arson                <NA>  2019-05-22  7:00            UNP
## 21355                 <NA>                <NA>  2019-06-02 15:40            UNP
##       det_agent  discovered_date discovered_size    reported_date
## 12962        RA  2014-05-24  7:28              NA  2014-05-24  7:32
## 13717        FG 2015-05-18 10:48              NA 2015-05-18 10:50
## 18213       HAC 2017-09-16 16:45              NA 2017-09-16 16:45
## 20474       310             <NA>              NA 2020-06-28 10:09
## 20778       LFS             <NA>              NA  2019-05-22  7:14
## 21355       PUB 2019-06-02 15:50              NA 2019-06-02 15:50
##       dispatched_resource     dispatch_date start_for_fire_date
## 12962               HAC  2014-05-24  8:14     2014-05-24  8:29
## 13717         FPD Staff 2015-05-18 11:00     2015-05-18 11:00
```

```
## 18213                    HAC 2017-09-16 16:45      2017-09-16 16:45
## 20474                    HAC 2020-06-28 14:30      2020-06-28 15:15
## 20778                    HAC  2019-05-22 9:27       2019-05-22 9:49
## 21355                    HAC 2019-06-02 15:59      2019-06-02 16:00
##       assessment_resource assessment_datetime assessment_hectares
## 12962           IA Forces    2014-05-24 10:35                0.20
## 13717               Other    2015-05-18 11:16                0.02
## 18213           IA Forces    2017-09-16 16:45                0.10
## 20474           IA Forces    2020-06-28 18:00                0.01
## 20778           IA Forces    2019-05-22 10:06                0.01
## 21355           IA Forces    2019-06-02 16:05                0.04
##       fire_spread_rate fire_type fire_position_on_slope
## 12962               -1    Ground                   Flat
## 13717               -1   Surface                   Flat
## 18213               -1   Surface                 Bottom
## 20474               -1   Surface                 Bottom
## 20778               -1    Ground                   Flat
## 21355               -1   Surface                   Flat
##       weather_conditions_over_fire temperature relative_humidity wind_direction
## 12962                  Rainshowers        10.5                73              N
## 13717                        Clear        20.0                22             SE
## 18213                        Clear        15.6                32              S
## 20474                       Cloudy        11.0                75              W
## 20778                        Clear        17.0                30              E
## 21355                       Cloudy        18.0                41              W
##       wind_speed fuel_type initial_action_by ia_arrival_at_fire_date
## 12962          1        S1          Industry                    <NA>
## 13717         10        D1         FPD Staff        2015-05-18 11:14
## 18213         20       O1b               HAC        2017-09-16 16:51
## 20474          2      <NA>               HAC        2020-06-28 17:53
## 20778          5        M2               HAC        2019-05-22 10:06
## 21355         15        C2               HAC        2019-06-02 16:05
##                ia_access fire_fighting_start_date fire_fighting_start_size
## 12962               <NA>                     <NA>                       NA
## 13717               <NA>         2015-05-18 11:16                     0.02
## 18213 Conventional   R/W         2017-09-16 16:57                     0.10
## 20474             Ground         2020-06-28 18:00                     0.01
## 20778             Ground         2019-05-22 10:13                     0.01
## 21355 Conventional   R/W         2019-06-02 16:10                     1.00
##       bucketing_on_fire distance_from_water_source first_bucket_drop_date
## 12962              <NA>                         NA                   <NA>
## 13717                 Y                        0.2       2015-05-18 11:16
## 18213                 N                         NA                   <NA>
## 20474                 N                         NA                   <NA>
## 20778                 N                         NA                   <NA>
## 21355                 Y                        0.1       2019-06-02 16:25
##              bh_fs_date bh_hectares         uc_fs_date uc_hectares to_fs_date
## 12962 2014-05-24 10:35        0.20 2014-05-24 11:40        0.20       <NA>
## 13717 2015-05-18 11:16        0.02 2015-05-18 11:50        0.02       <NA>
## 18213 2017-09-16 17:09        0.10 2017-09-16 17:55        0.20       <NA>
## 20474 2020-06-28 18:00        0.01 2020-06-28 18:00        0.01       <NA>
## 20778 2019-05-22 10:06        0.01 2019-05-22 10:06        0.01       <NA>
## 21355 2019-06-02 16:05        0.04 2019-06-02 19:21        0.04       <NA>
##       to_hectares         ex_fs_date ex_hectares
```

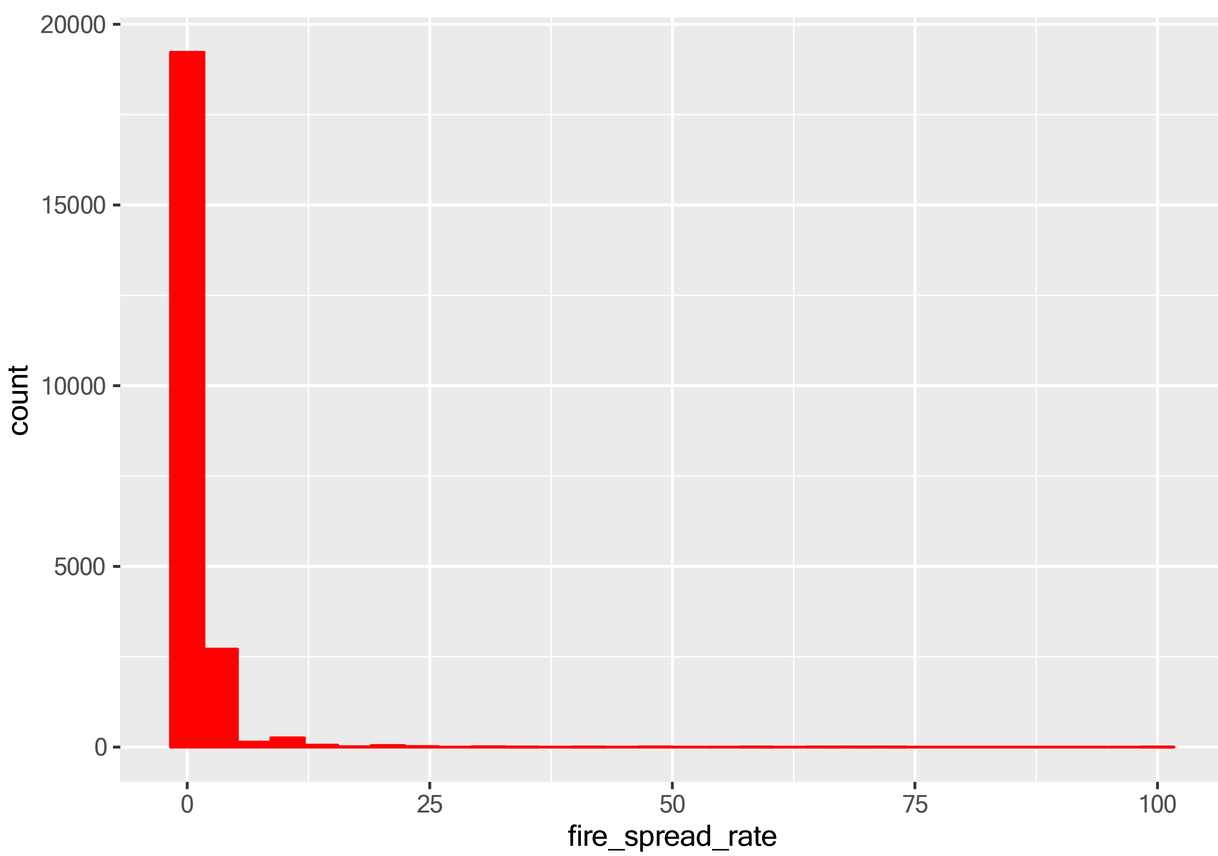
```
## 12962               NA  2014-05-24  14:31            0.20
## 13717               NA  2015-05-18  13:00            0.02
## 18213               NA  2017-09-17  10:50            0.10
## 20474               NA  2020-06-28  19:56            0.01
## 20778               NA  2019-05-22  10:25            0.01
## 21355               NA  2019-06-02  19:50            0.04
```

Upon visual inspection there does not seem to be any pattern related to the fire_spread_rate being -1. Based on this, and the definition provided in the data dictionary, with fire_spread_rate being 'The rate of spread of the wildfire at the time of initial assessment, capture in metres per minute', we felt it was safe to remove these rows, as this is most likely an error with these entries. For the fire to have a negative spread rate, would mean that the fire is retreating instead of spreading, and given that this rate of spread is a measure of how fast the fire moves from a point of origin, this seems counter intuitive to how forest fires work. Given more time, we could reach out to the providers of the data, to try to clarify this area, but for the time being, and since there are only 6 data points, we will remove them.

After we remove the rows with a fire spread rate of -1, we can plot the data for a preliminary visualization.

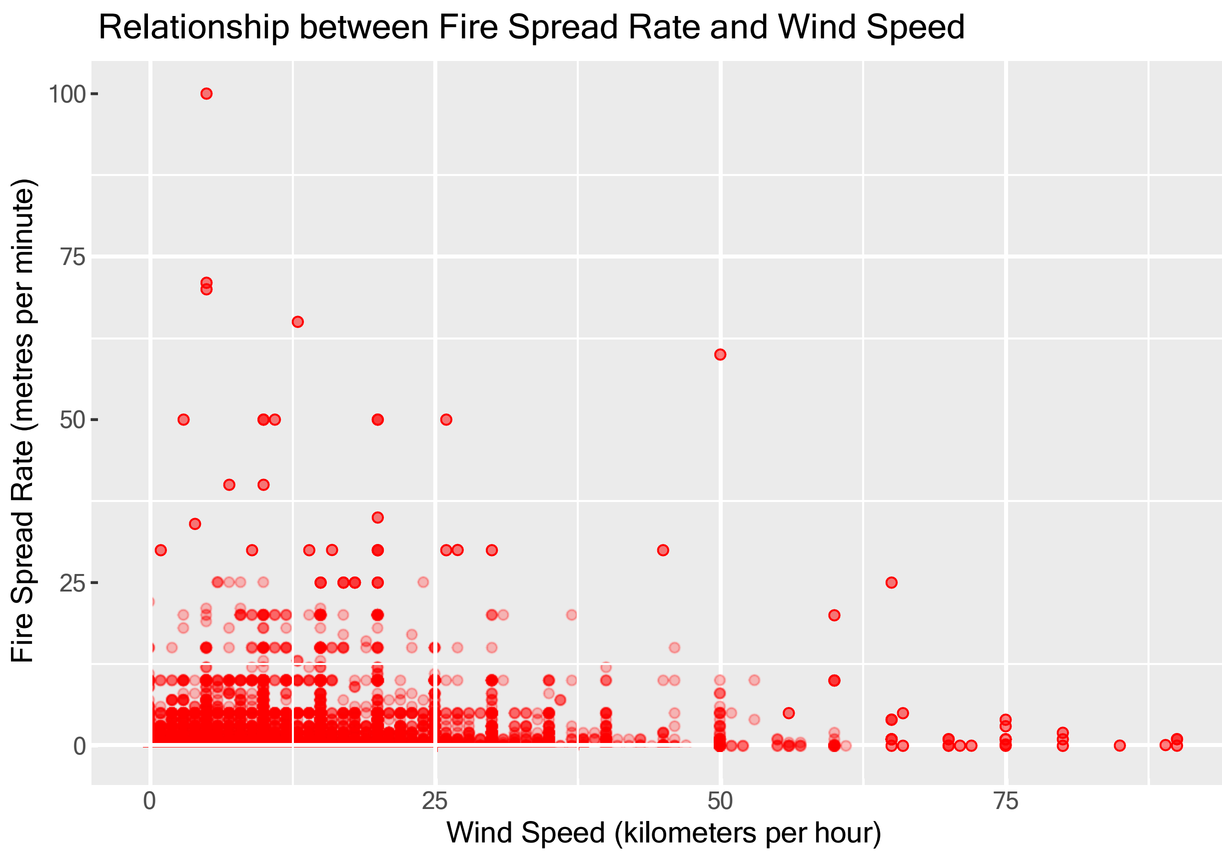
```r
fire_clean_no_neg=fire_clean[fire_clean['fire_spread_rate']>=0,]
ggplot(fire_clean_no_neg, aes(x = fire_spread_rate)) +
  geom_histogram(color='red',fill='red')
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Based on our histogram, we can see that the data for fire_spread_rate is heavily skewed.

```r
fire_clean_no_neg=fire_clean[fire_clean['fire_spread_rate']>=0,]
ggplot(fire_clean_no_neg, aes(x = wind_speed, y = fire_spread_rate)) +
  geom_point(color='red',alpha = 0.5) +
  labs(title = "Relationship between Fire Spread Rate and Wind Speed",
       x = "Wind Speed (kilometers per hour)",    # Replace with actual units if known
       y = "Fire Spread Rate (metres per minute)")
```



Initial inspection of the scatterplot is difficult to arrive to any meaningful conclusion without further analysis.

Additionally, we will look at the correlation coefficient between fire_spread_rate and wind_speed.

```r
fire_corr=cor(fire_clean_no_neg,use="pairwise.complete.obs")
print(fire_corr)
```

```
##                  wind_speed fire_spread_rate
## wind_speed        1.0000000        0.1346716
## fire_spread_rate  0.1346716        1.0000000
```

Based on our output we can see a very weak positive relationship between fire spread rate and wind speed.

```r
fire_no_neg_model=lm(fire_spread_rate ~ wind_speed, data = fire_clean_no_neg)

summary(fire_no_neg_model)
```

```
##
## Call:
## lm(formula = fire_spread_rate ~ wind_speed, data = fire_clean_no_neg)
##
## Residuals:
##    Min      1Q  Median     3Q    Max
##  -4.263  -0.863  -0.597   0.054  99.261
##
## Coefficients:
##              Estimate  Std. Error  t value  Pr(>|t|)
##  (Intercept)  0.531266   0.024815    21.41    <2e-16 ***
## wind_speed  0.041468   0.002035    20.38    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.571 on 22478 degrees of freedom
## Multiple R-squared:   0.01814,     Adjusted R-squared:   0.01809
## F-statistic: 415.2 on 1 and 22478 DF,    p-value: < 2.2e-16
```

Intercept: The expected value of fire_spread_rate when wind_speed is zero is 0.53130. As our p-value is <0.05, this indicates that we can reject the H_0 that β_0=0. Therefore, we accept the H_1 that β_0/= 0 and conclude that the intercept is statistically significant.

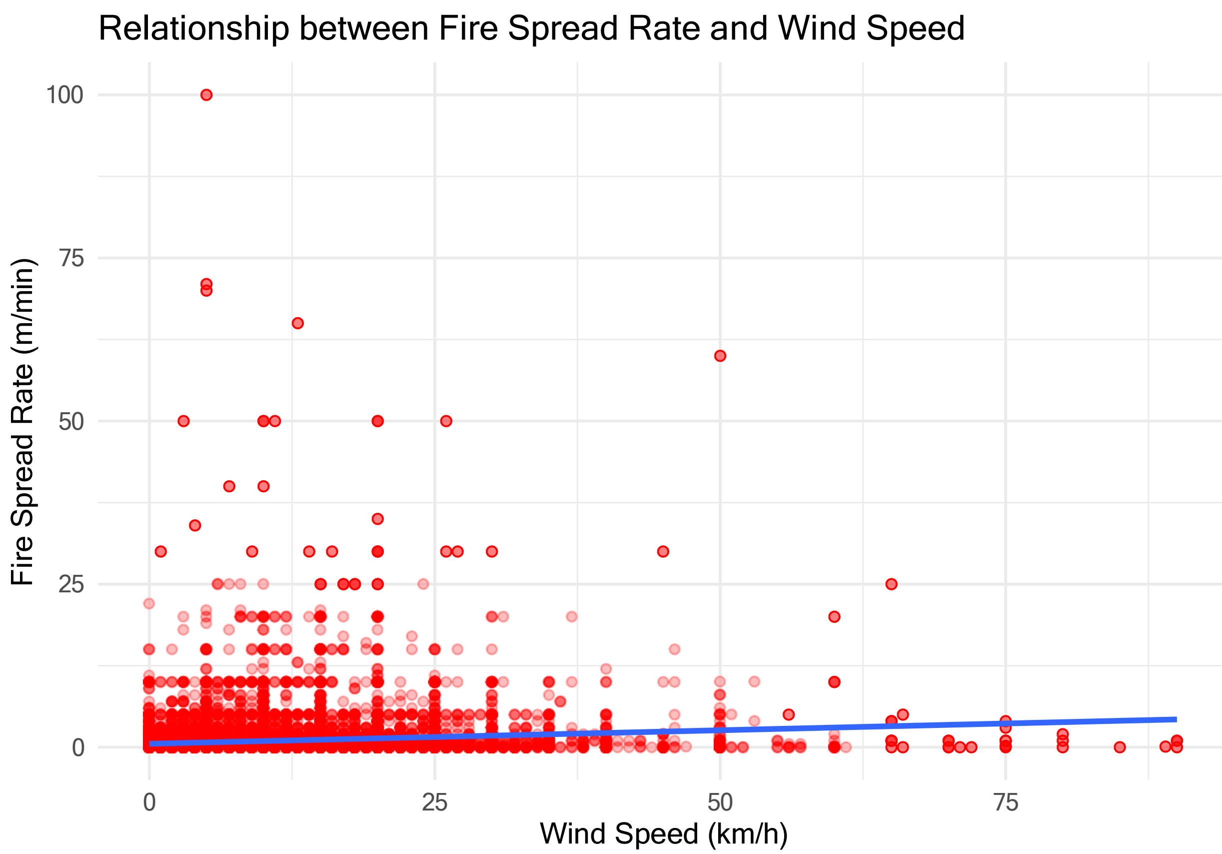
Slope: For each additional kilometer per hour in wind_speed, the fire_spread_rate is expected to increase by approximately 0.04150 meters per minute. As our p-value is <0.05, this indicates that we can reject the H_0 that β_1=0. Therefore, we accept the H_1 that β_1/= 0 and conclude that there is a significant relationship between wind_speed and fire_spread_rate.

Based on our output table, the equation for our model can be written out as, $fire\_spread\_rate = 0.53130 + (0.04150 * wind\_speed)$

Our R-squared value indicates that approximately 1.81% of the variance in fire_spread_rate is explained by wind_speed. This low value suggests that there are other factors affecting fire spread that are not included in our model.

We can plot this model using the following code:

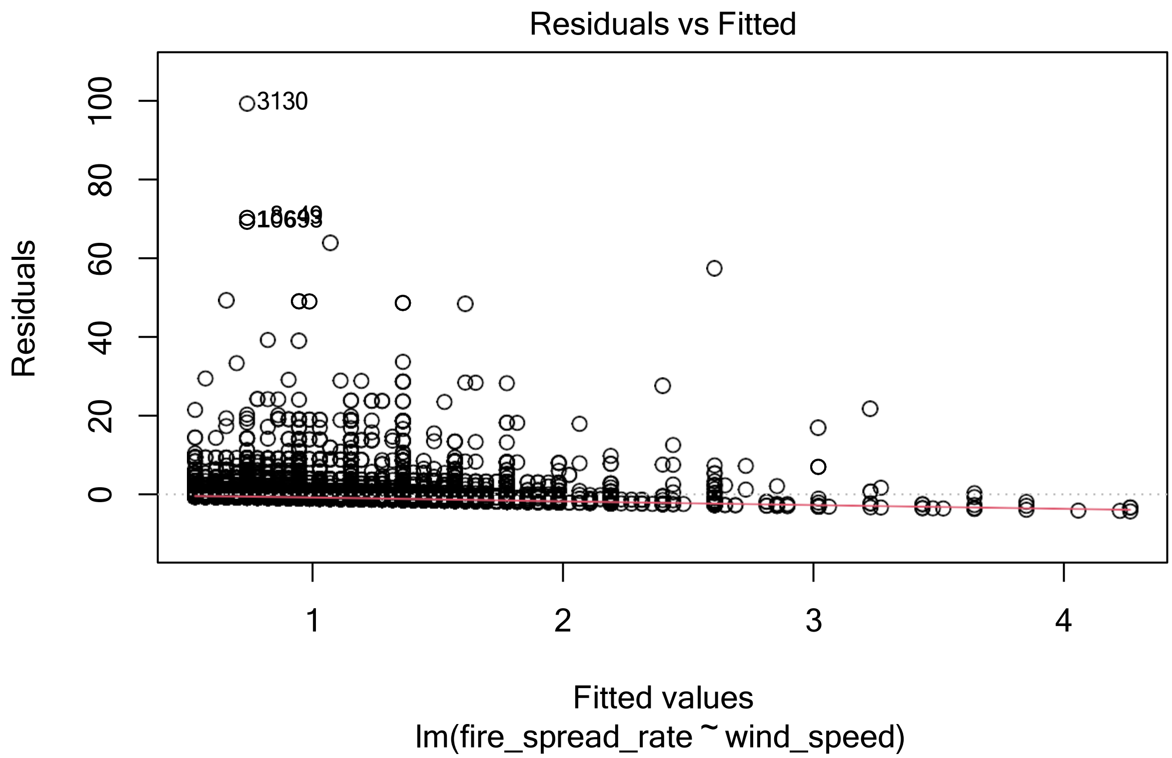
```
ggplot(fire_clean_no_neg, aes(x = wind_speed, y = fire_spread_rate)) +
  geom_point(alpha = 0.5,color='red') +
  stat_smooth(method = "lm", formula = y~x) +   # Add regression line
  labs(title = "Relationship between Fire Spread Rate and Wind Speed",
       x = "Wind Speed (km/h)",
       y = "Fire Spread Rate (m/min)") +
  theme_minimal()
```



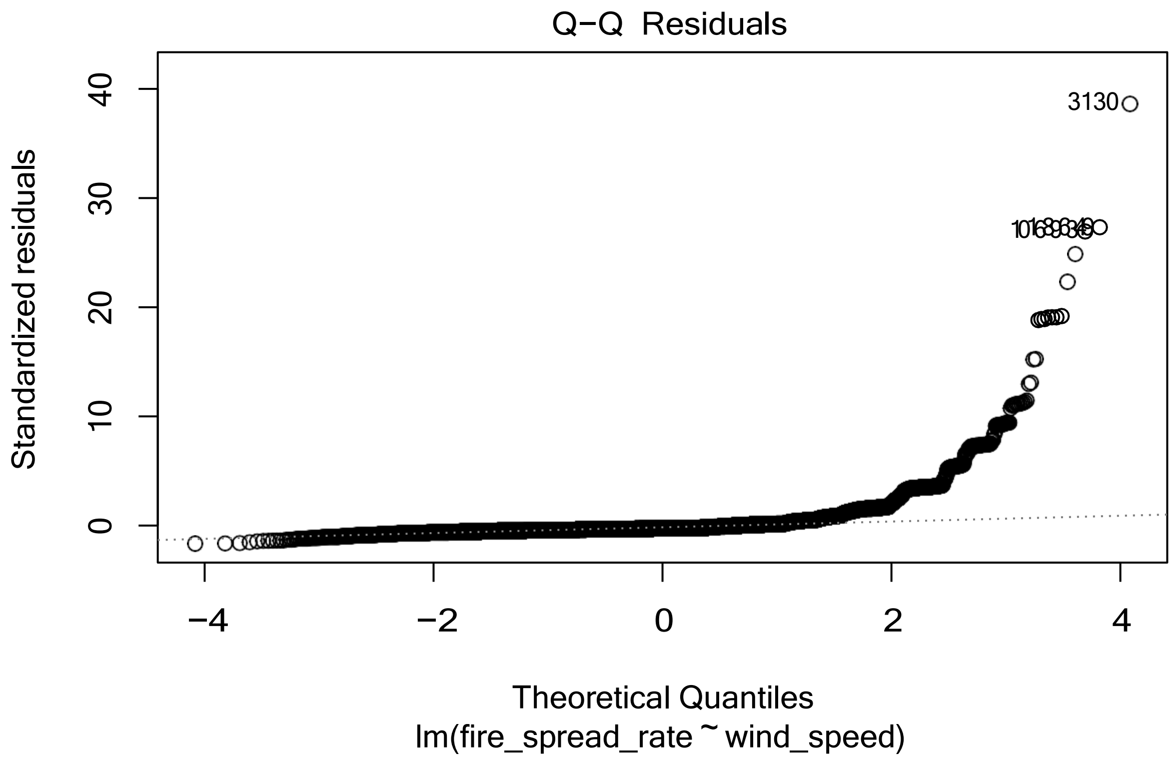
From our output, we can see a weak relationship between wind_speed and fire_spread_rate. as suggested by our correlation statistics.

To determine whether the assumptions of independency and normality are met, we can plot the residual and QQ plots.

```
#Residual plot
plot(fire_no_neg_model, which =1)
```




```r
#QQ-plot
plot(fire_no_neg_model, which =2)
```



For our residual plot we see a large cluster of points on the upper left side of the figure, as well as some outliers. For our QQ plot, we see that the points stray off far from the line towards the right side of the figure. Based on these observation, we can suggest that both assumptions of independency and normality of residuals fails.

By failing both of these assumptions, it suggests issues in our model that can lead to unreliable results. Some potential solutions are to transform the data, or to include interaction terms.

Question 2 - What is the Relationship Between Temperature and Fire Spread Rate

In this project, we explored the relationship between temperature and fire spread rate in Canada. We visualized the distribution of fire spread rate and temperature, calculated the correlation coefficient, performed a linear regression analysis, and conducted a hypothesis test to determine whether the observed relationship is statistically significant. Additionally, we created a geospatial representation of fire spread rate along with temperature to better understand the spatial patterns and relationships between these variables.

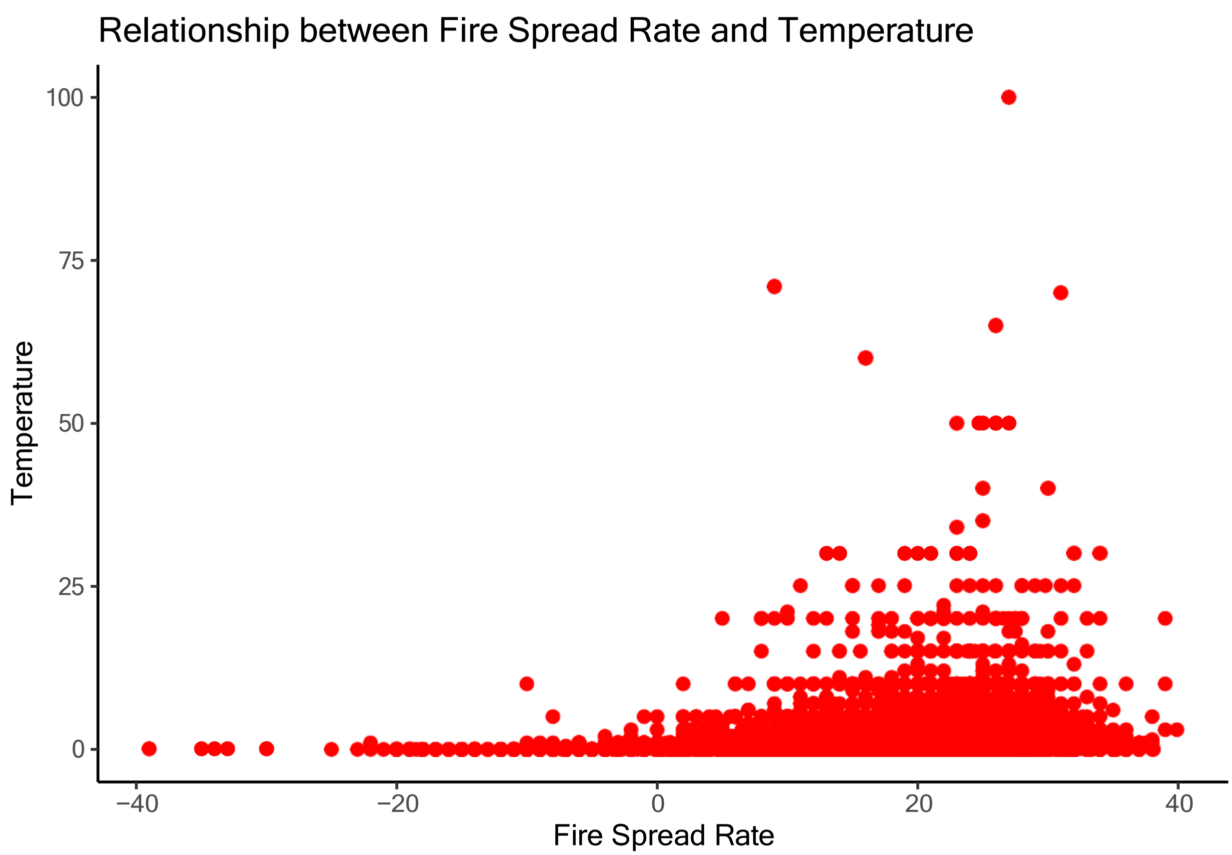
The results of the hypothesis test indicate whether there is a statistically significant relationship between temperature and fire spread rate.

```r
data=fire_data
```

```r
# Filter out rows where Fire Spread Rate is negative
data<- data %>%
  filter(fire_spread_rate >= 0)
# Create a scatter plot
ggplot(data, aes(x = temperature, y = fire_spread_rate)) +
  geom_point(color = "red", size = 2) +
```

```r
  #geom_smooth(method = "lm", se = FALSE, color = "red")

 labs(title = "Relationship between Fire Spread Rate and Temperature",
       x = "Fire Spread Rate",
       y = "Temperature") +
  theme_classic()
```

Warning: Removed 80 rows containing missing values or values outside the scale range
('geom_point()').



EDA

```r
# Check the structure of the data
str(data)
```

```
## 'data.frame':       22562 obs. of   50 variables:
##   $ fire_year                  : int   2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
##   $ fire_number                : chr   "PWF001" "EWF002" "EWF001" "EWF003" ...
##   $ fire_name                  : chr   NA NA NA NA ...
##   $ current_size               : num  0.1 0.2 0.5 0.01 0.1 0.2 0.01 0.01 0.2 0.6 ...
##   $ size_class                 : chr   "A" "B" "B" "A" ...
##   $ fire_location_latitude     : num  56.2 53.6 53.6 53.6 56.2 ...
```

```
##  $ fire_location_longitude  : num  -117 -116 -116 -116 -117 ...
##  $ fire_origin              : chr  "Private Land" "Provincial Land" "Provincial Land" "Provincial
##  $ general_cause_desc       : chr  "Resident" "Incendiary" "Incendiary" "Incendiary" ...
##  $ industry_identifier_desc : chr  NA NA NA NA ...
##  $ responsible_group_desc   : chr  "Resident" "Others (explain in remarks)" "Others (explain in
##  $ activity_class           : chr  "Grass" "Lighting Fires" "Lighting Fires" "Lighting Fires" ...
##  $ true_cause               : chr  "Permit Related" "Arson Suspected" "Arson Suspected" "Arson Su
##  $ fire_start_date          : chr  "2006-04-02 12:00" "2006-04-03 12:10" "2006-04-03 12:15" "2006
##  $ det_agent_type           : chr  "UNP" "UNP" "UNP" "UNP" ...
##  $ det_agent                : chr  "310" "310" "310" "PUB" ...
##  $ discovered_date          : chr  NA NA NA NA ...
##  $ discovered_size          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ reported_date            : chr  "2006-04-02 20:46" "2006-04-03 12:27" "2006-04-03 12:36" "2006
##  $ dispatched_resource      : chr  "FPD Staff" "FPD Staff" "FPD Staff" "FPD Staff" ...
##  $ dispatch_date            : chr  "2006-04-02 21:10" "2006-04-03 12:33" "2006-04-03 12:36"  "2006
##  $ start_for_fire_date      : chr  "2006-04-02 21:20" "2006-04-03 12:35" "2006-04-03 12:42"  "2006
##  $ assessment_resource      : chr  "IA Forces" "IA Forces" "IA Forces" "IA Forces" ...
##  $ assessment_datetime      : chr  "2006-04-02 22:00" "2006-04-03 13:20" "2006-04-03 13:23"  "2006
##  $ assessment_hectares      : num  0.01 0.2 0.5 0.01 0.1 0.2 0.01 0.01 0.2 0.6 ...
##  $ fire_spread_rate         : num  0 0 0 0 0.1 0 0 0 0 0 ...
##  $ fire_type                : chr  "Surface" "Surface" "Surface" "Surface" ...
##  $ fire_position_on_slope   : chr  "Flat" "Lower 1/3" "Bottom" "Flat" ...
##  $ weather_conditions_over_fire: chr  "Clear" "Clear" "Clear" "Clear" ...
##  $ temperature              : num  18 12 12 12 6 11 11 16 11 11 ...
##  $ relative_humidity        : int  10 22 22 22 37 32 25 17 35 44 ...
##  $ wind_direction           : chr  "SW" "SW" "SW" "SW" ...
##  $ wind_speed               : int  2 10 10 10 2 20 10 2 7 4 ...
##  $ fuel_type                : chr  "O1a" "O1a" "O1a" "O1b" ...
##  $ initial_action_by        : chr  "Land Owner" "Fire Department" "Fire Department" "Industry" ..
##  $ ia_arrival_at_fire_date  : chr  NA NA NA NA ...
##  $ ia_access                : chr  NA NA NA NA ...
##  $ fire_fighting_start_date : chr  NA NA NA NA ...
##  $ fire_fighting_start_size : num  NA NA NA NA NA NA NA 0.01 NA 0.6 ...
##  $ bucketing_on_fire        : chr  NA NA NA NA ...
##  $ distance_from_water_source : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ first_bucket_drop_date   : chr  NA NA NA NA ...
##  $ bh_fs_date               : chr  "2006-04-02 22:00" "2006-04-03 13:20" "2006-04-03  13:23"  "2006
##  $ bh_hectares              : num  0.01 0.2 0.5 0.01 0.1 0.2 0.01 0.01 0.2 0.6 ...
##  $ uc_fs_date               : chr  "2006-04-02 22:00" "2006-04-03 13:20" "2006-04-03  13:23"  "2006
##  $ uc_hectares              : num  0.01 0.2 0.5 0.01 0.1 0.2 0.01 0.01 0.2 0.6 ...
##  $ to_fs_date               : chr  NA NA NA NA ...
##  $ to_hectares              : num  NA NA NA NA 0.1 NA NA 0.01 0.2 NA ...
##  $ ex_fs_date               : chr  "2006-04-03 10:20" "2006-04-03 14:00" "2006-04-03  15:00"  "2006
##  $ ex_hectares              : num  0.1 0.2 0.5 0.01 0.1 0.2 0.01 0.01 0.2 0.6 ...
```

```
# View unique values in Temperature and FireSpreadRate
unique(data$temperature)
```

```
##  [1]  18.0  12.0   6.0  11.0  16.0  28.0  26.0  25.0  35.0  15.0  10.0  13.2
## [13]  27.0  20.0  17.0  22.0  24.0  23.0  21.0  29.0  19.0  21.4  14.5  14.0
## [25]  33.0  31.0  32.0   4.0  24.5  30.0  27.5  22.5  21.5   6.5   9.0  13.0
## [37]  25.5   2.0  30.6  17.5   5.0  12.5  22.4  36.0  26.4  18.7  16.1  21.6
## [49]  16.5  19.5   7.0  24.6  -0.6  20.5  23.5  24.3  22.1  18.4  23.7  26.5
## [61]   8.0  15.6  22.8  12.8  13.5  29.5  14.2   8.3  11.5   8.5  13.3  17.6
```

```
##  [73]  26.2  19.6   9.4   7.5  16.7  21.2   5.4   3.0   1.0  18.5  23.2  16.3
##  [85]  24.8  26.7  19.8  31.5  25.6  28.2  28.7  20.2  22.7  17.3  30.8  24.7
##  [97]  23.6  17.4  22.3  28.6 -18.0   0.1  -1.0  30.9  -8.0  -7.0  11.2   0.0
## [109]  19.4 -10.0  -5.0   6.2  11.7  -2.0  14.4  21.1  25.7  26.6  10.8  15.3
## [121]   0.2  -7.3  -4.0   7.6   8.9 -11.0  28.5  13.6  30.7  -6.0 -15.0  -3.0
## [133]   5.5  26.1   9.5  14.9  15.5  13.4  16.9  24.4  17.8  19.9  15.4  26.8
## [145]  25.4  21.8  18.2  27.2  -3.5 -12.0 -14.0 -13.0  20.9   2.5  14.6  12.1
## [157]  16.2  12.2  19.2  19.1  10.6  -9.0   8.2   6.6  20.6  17.7   3.2  20.7
## [169]  18.6   8.6  20.1  15.2  11.8  22.6  28.3   3.5  -2.7   6.1   4.8  13.8
## [181]  10.5   2.4  14.3  28.9  34.0 -21.0  26.3 -18.5   7.4  22.2  11.9  24.1
## [193]  14.8  20.3  18.3  29.1   8.7   7.2   5.1  27.8  15.7 -20.0  31.6   2.6
## [205]   5.7  13.7  12.4  27.4 -33.0   8.1   9.1    NA -25.0   8.8  30.5  14.1
## [217]  19.7   6.4   4.5  10.4  24.9  24.2 -17.0  10.2  13.9  18.1  23.9 -22.0
## [229]  10.3 -19.0   8.4 -16.0 -30.0  14.7  23.4  20.4  19.3  32.1  16.6  12.3
## [241]  18.8  25.8  37.0   1.5  16.4  -1.5  25.2   5.6  29.4  21.3  27.6  28.1
## [253]  27.1   9.3  18.9  12.7   6.3  27.7  32.5  23.8  33.2  25.3  20.8  17.9
## [265]  23.3  27.9  29.3  15.9  21.9  17.1  21.7  28.4  17.2  15.1  16.8   0.5
## [277]  32.4  29.8  30.4  39.0 -35.0 -34.0   2.3  13.1  28.8   3.3   6.7  23.1
## [289]   2.2  29.2  11.6   5.3  -2.5  29.6  33.4  27.3  11.1   1.7  31.8  22.9
## [301]  15.8   9.7  29.7   9.2  31.7  -3.4  35.4  31.2   7.3  30.1   1.3  12.6
## [313]   5.8  25.9   1.9  31.3   6.8   0.7   3.6   3.7  26.9 -23.0  25.1  38.0
## [325]   0.3  34.5  35.3  29.9  39.9   9.9  38.1   1.6  11.3  32.6   5.9  32.7
## [337]  37.5   4.7  12.9 -39.0  -0.2  10.1   7.9   9.8
```

```r
unique(data$fire_spread_rate)
```

```
##  [1]   0.0   0.1   1.0  12.0   0.5   1.5   3.0   2.0   5.0  10.0   0.2  35.0
## [13]   4.0  30.0   0.4  50.0   0.9   8.0  20.0   3.5   7.0   0.3   6.0  11.0
## [25]  17.0   9.0  25.0  18.0  40.0   2.5 100.0   0.7  15.0   0.8   1.2   1.1
## [37]   1.8   1.9  13.0   4.5   1.7   8.5  70.0   0.6   2.2   4.9  16.0  65.0
## [49]  21.0   5.7  34.0  71.0  19.0  60.0  22.0   5.5
```

We will also look at a summary of the data

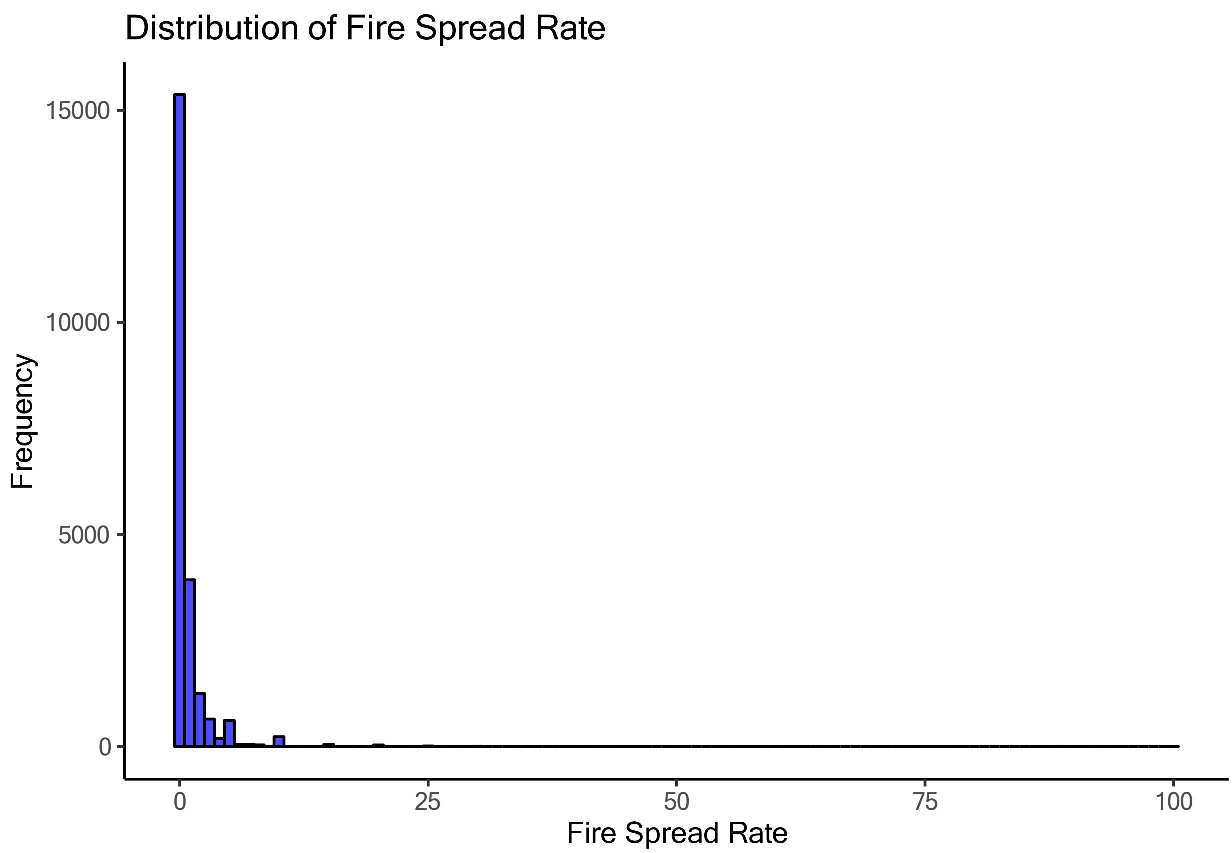
```r
summary(data$temperature)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -39.00   14.00   19.00   17.85   23.00   39.90      80
```

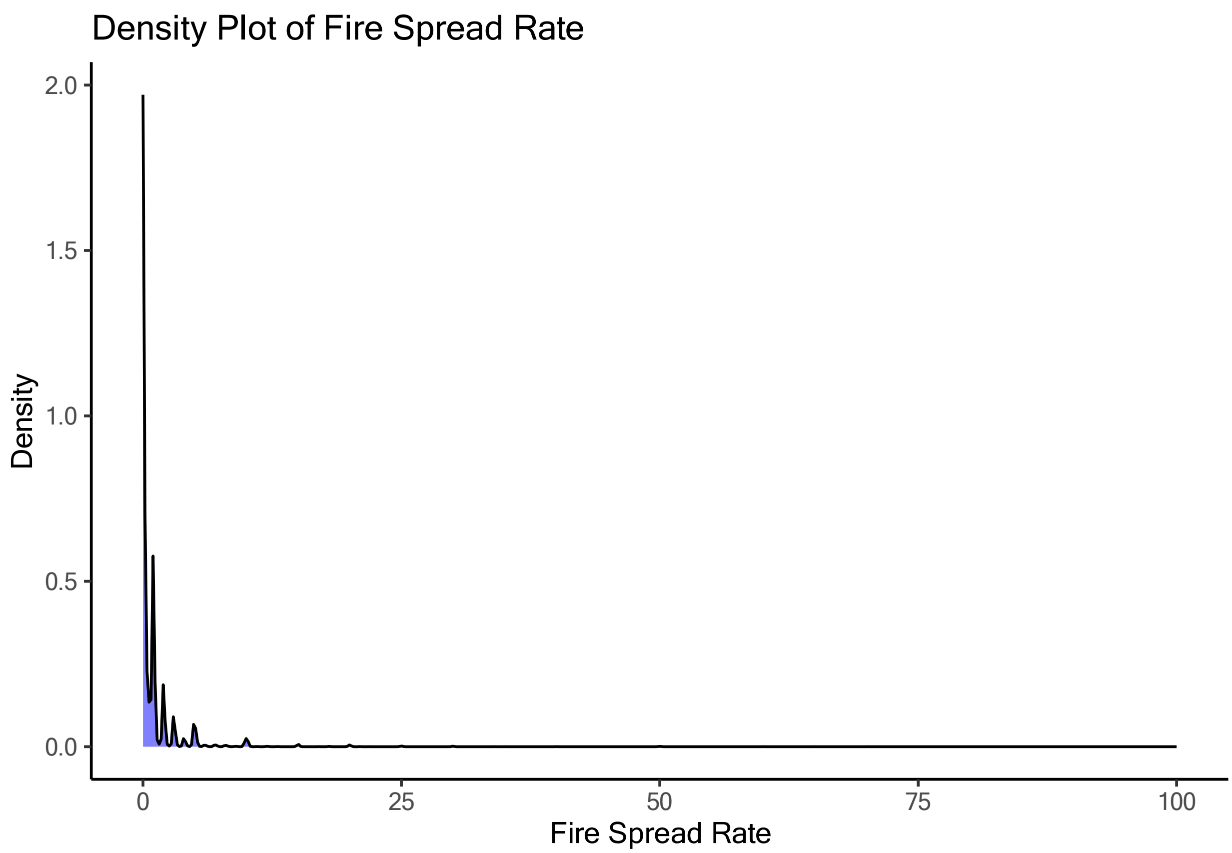
```r
summary(data$fire_spread_rate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.0000   0.8962   1.0000 100.0000
```

```r
# Plot a histogram
ggplot(data, aes(x = fire_spread_rate)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Fire Spread Rate",
       x = "Fire Spread Rate",
       y = "Frequency") +
  theme_classic()
```




```r
# Alternatively, plot a density plot
ggplot(data, aes(x = fire_spread_rate)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot of Fire Spread Rate",
       x = "Fire Spread Rate",
       y = "Density") +
  theme_classic()
```




```r
# Calculate skewness
spread_rate_skewness <- skewness(data$fire_spread_rate)
print(paste("Skewness of Fire Spread Rate:", spread_rate_skewness))
```

```
## [1] "Skewness of Fire Spread Rate: 11.2632058482397"
```

The value 11.22 suggests that the distribution of fire spread rates is heavily skewed to the right, meaning that most of the fire spread rates are relatively low, but there are a few extremely high values (outliers) that pull the tail of the distribution to the right.

Before performing a regression analysis, it is important to investigate the correlation between both variables.

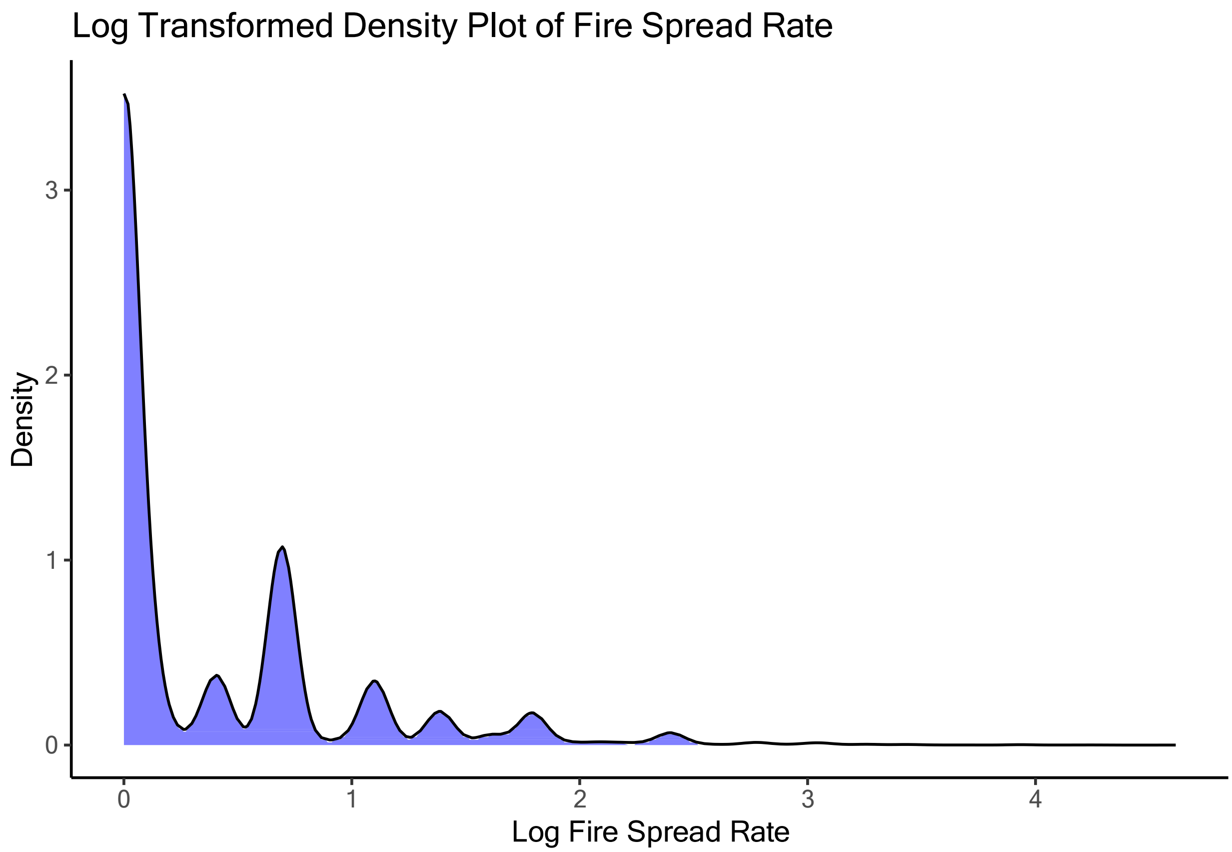
To normalize the distribution of the fire spread rate we can attempt a log transformation

```r
# Log Transformation
data$log_fire_spread_rate <- log(data$fire_spread_rate + 1)

log_spread_rate_skewness <- skewness(data$log_fire_spread_rate)
print(paste("Skewness of Log Transformed Fire Spread Rate:", log_spread_rate_skewness))
```

```
## [1] "Skewness of Log Transformed Fire Spread Rate: 1.92868885555048"
```

```r
ggplot(data, aes(x = log_fire_spread_rate)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Log Transformed Density Plot of Fire Spread Rate",
```

```
       x = "Log Fire Spread Rate",
       y = "Density") +
  theme_classic()
```



Following this, we will calculate Pearson correlation between log-transformed fire spread rate and temperature

```
correlation_log <- cor(data$log_fire_spread_rate, data$temperature, use = "complete.obs")
print(paste("Pearson correlation coefficient (Log Fire Spread Rate and Temperature):", correlation_log)
```

```
## [1] "Pearson correlation coefficient (Log Fire Spread Rate and Temperature): 0.250089700065433"
```

We will now rechecking correlation fire spread rate and temperature to see if the relationship has improved now.

The regression model can be coded as:

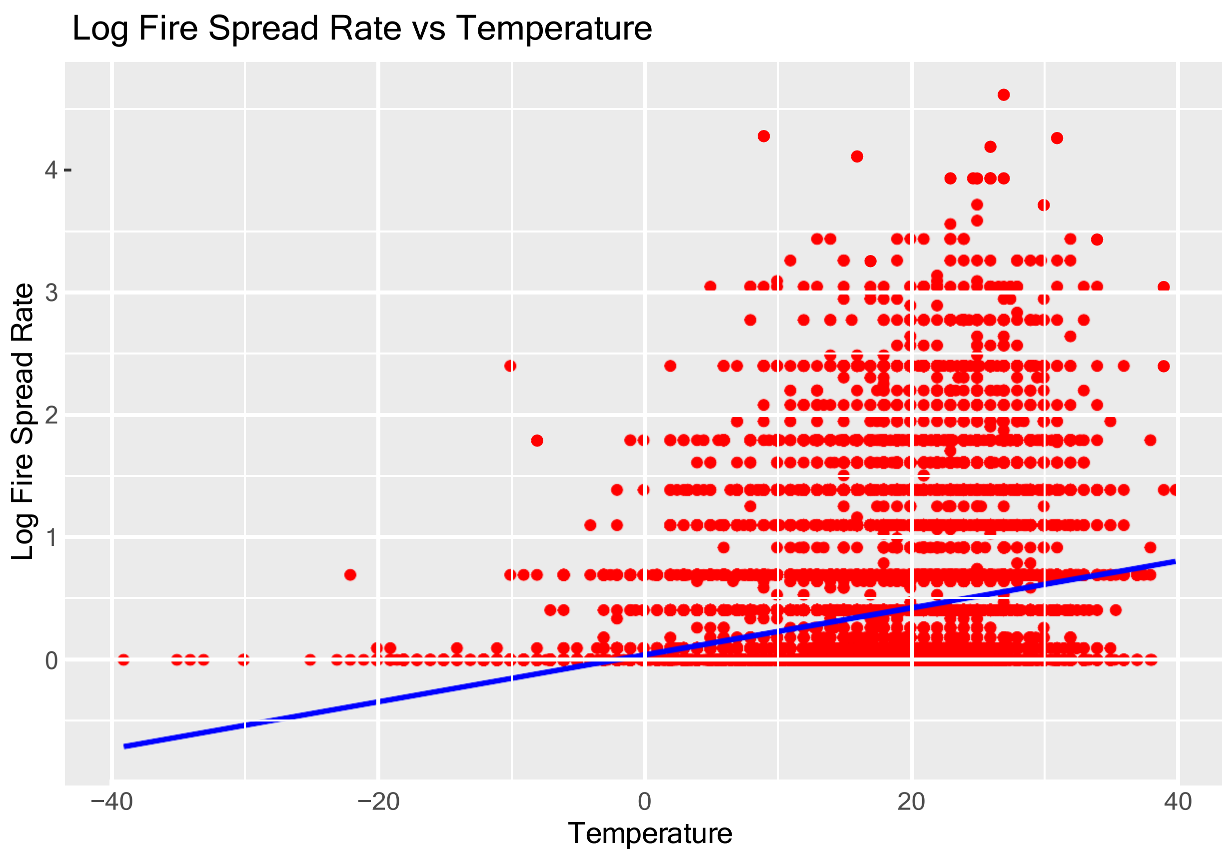
```
# Fit a linear regression model
log_model <- lm(log_fire_spread_rate ~ temperature, data = data)
summary(log_model)
```

```
##
## Call:
## lm(formula = log_fire_spread_rate ~ temperature, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.7690  -0.3839  -0.2119   0.2326   4.0652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0389989  0.0096034   4.061  4.9e-05 ***
## temperature 0.0191612  0.0004948  38.727  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 22480 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.06254,    Adjusted R-squared:  0.0625
## F-statistic:  1500 on 1 and 22480 DF,  p-value: < 2.2e-16
```

```r
ggplot(data, aes(x = temperature, y = log_fire_spread_rate)) +
  geom_point(color = "red") +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Log Fire Spread Rate vs Temperature",
       x = "Temperature",
       y = "Log Fire Spread Rate")
```

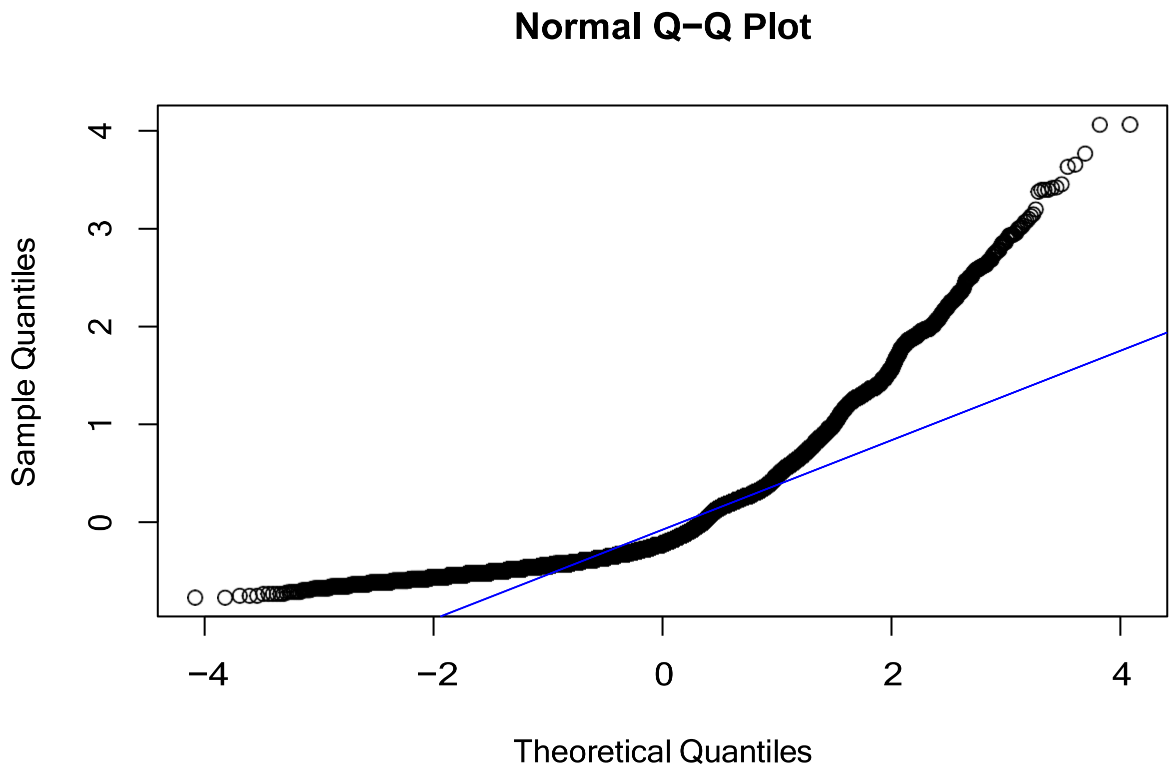
```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



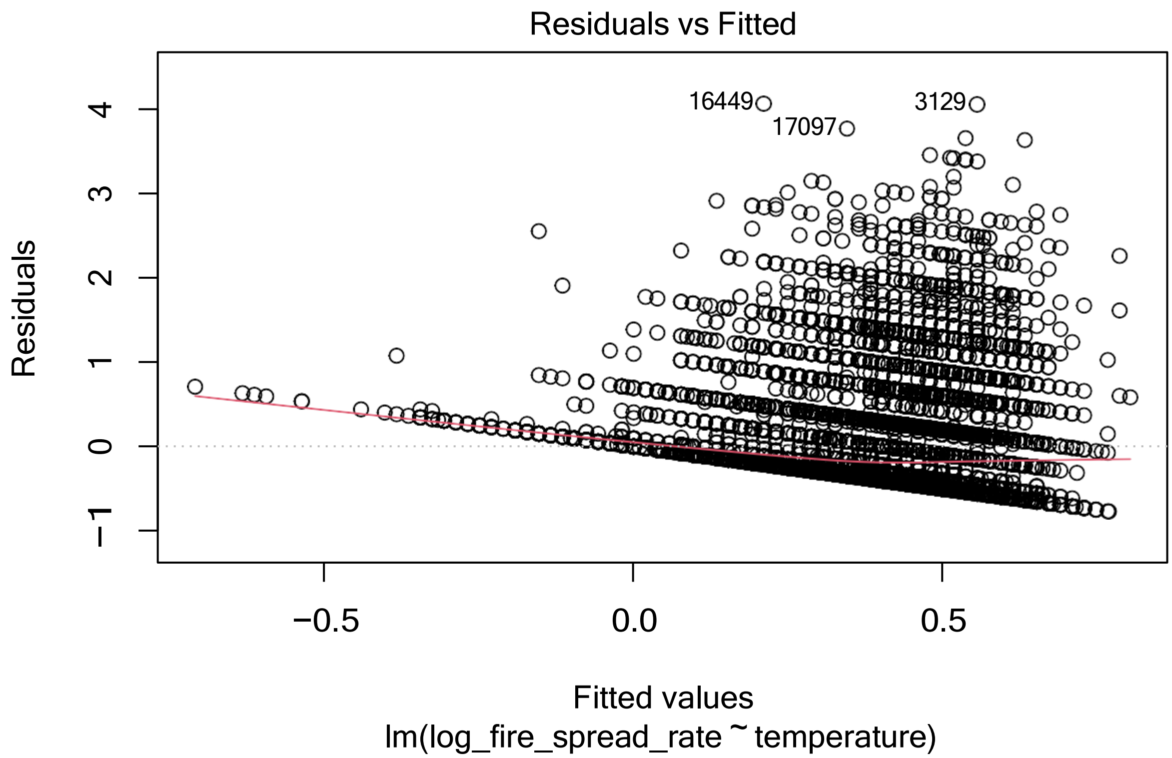
The graph shows the relationship between temperature and the rate of fire spread. The red dots represent data points where the x-axis represents the temperature and the y-axis represents the log of the fire spread rate. The blue line represents a line of best fit, indicating that there is a positive correlation between the two variables. This means that as temperature increases, the fire spread rate also increases, and this is illustrated by the upward trend of the blue line.

Q-Q Plot and Residuals Plot to check the normality

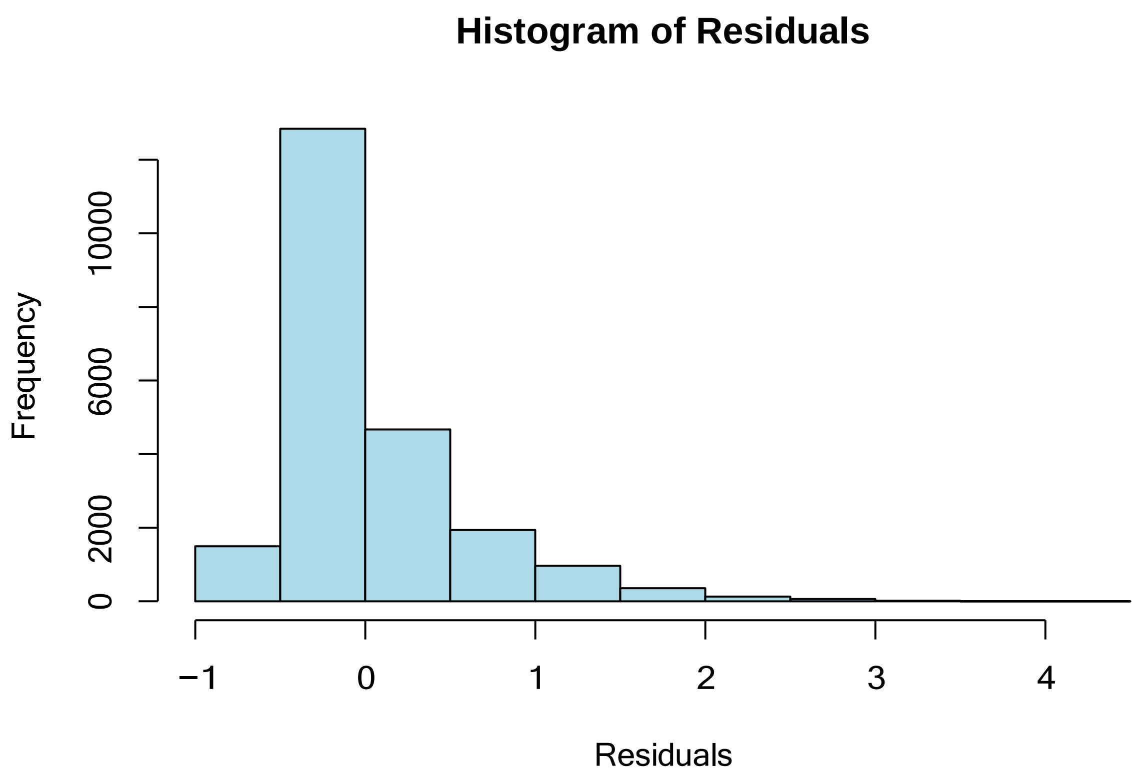
```
# Q-Q plot to check normality of residuals
qqnorm(residuals(log_model))
qqline(residuals(log_model), col = "blue")
```




```r
# Residuals vs Fitted plot
plot(log_model, which = 1)
```




```
# Histogram of residuals
hist(residuals(log_model), main = "Histogram of Residuals", xlab = "Residuals", col = "lightblue")
```



Residual plot: This plot suggests that the linear model is not a good fit for the data, and alternative models might be considered.

Q-Q Residual: This plot indicates that the data is not normally distributed because it deviates from a straight line. The points are curved and have a few outliers. The data is skewed to the right as it deviates from the straight line on the right-hand side.

```
# Get model summary and extract statistics
log_model_summary <- summary(log_model)
r_squared_log <- log_model_summary$r.squared
adj_r_squared_log <- log_model_summary$adj.r.squared
p_value_log <- log_model_summary$coefficients[2, 4]

print(paste("R-squared (log model):", r_squared_log))
```

```
## [1] "R-squared (log model): 0.0625448580788178"
```

```
print(paste("Adjusted R-squared (log model):", adj_r_squared_log))
```

```
## [1] "Adjusted R-squared (log model): 0.062503156337629"
```

```
print(paste("p-value for temperature (log model):", p_value_log))
```

```
## [1] "p-value for temperature (log model): 1.1251791730511e-317"
```

To avoid the complexity by temperature variable Adjusted R-squared -the relationship between the temperature and fire spread rate is weak, meaning temperature does not explain much of the variation in fire spread rate.P Value suggests that the temperature is likely statistically significant, meaning it has a real effect on the dependent variable.

To check this since the p-value only reflects the significance of the relationship, it's also worth considering if other variables (like humidity , wind speed , vegetation_type

```r
# Extra work--Consider including other relevant features
data$humidity <- data$relative_humidity
data$wind_speed <- data$wind_speed
data$vegetation_type <- data$fuel_type

# Update the model to include additional features
model <- lm(fire_spread_rate ~ temperature + humidity + wind_speed + vegetation_type, data  = data)



# Summarize the updated model
summary(model)
```

```
##
## Call:
## lm(formula = fire_spread_rate ~ temperature + humidity + wind_speed +
##     vegetation_type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.849  -1.028  -0.475   0.221 98.075
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.241233   0.161542   7.684 1.63e-14 ***
## temperature           0.031667   0.003164  10.009  < 2e-16 ***
## humidity             -0.017181   0.001239 -13.866  < 2e-16 ***
## wind_speed            0.044757   0.002431  18.413  < 2e-16 ***
## vegetation_typeC2     0.223939   0.121777   1.839 0.065942 .
## vegetation_typeC3    -0.246459   0.156810  -1.572 0.116036
## vegetation_typeC4    -0.443133   0.259062  -1.711 0.087185 .
## vegetation_typeC7    -0.845929   0.621074  -1.362 0.173202
## vegetation_typeD1    -0.759180   0.176130  -4.310 1.64e-05 ***
## vegetation_typeM1    -0.559852   0.160891  -3.480 0.000503 ***
## vegetation_typeM2    -0.699145   0.130445  -5.360 8.44e-08 ***
## vegetation_typeM3    -1.718422   1.577507  -1.089 0.276024
## vegetation_typeM4    -0.739332   2.726585  -0.271 0.786274
## vegetation_typeO1a   -0.760666   0.127503  -5.966 2.48e-09 ***
## vegetation_typeO1b   -0.719319   0.134277  -5.357 8.57e-08 ***
## vegetation_typeS1    -0.892669   0.173143  -5.156 2.55e-07 ***
## vegetation_typeS2    -1.039302   0.174980  -5.940 2.91e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.724 on 17918 degrees of freedom
##   (4627 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.07169,   Adjusted R-squared:  0.07086
## F-statistic: 86.48 on 16 and 17918 DF,  p-value: < 2.2e-16
```

This thing concludes that not only one factor is responsible for the fire spread these other factors combiningly also affect that.

Creating a new column according to the size_class to represent that in the visualization as High or low spread in the particular area.

```
# Create SpreadCategory based on size_class
data$SpreadCategory <- ifelse(data$size_class %in% c("D", "E"), "High", "Low")

#head(data) #masking output here to avoid clutter, as we have seen this before, want to leave it in to
```

Geospatial Visualization: Map the fire occurrences geographically with spread rate and temperature as visual layers

```
#For the following code, the output is an HTML output, therefore exporting it in a PDF format is not po

#data <- data[!is.na(data$fire_location_longitude) & !is.na(data$fire_location_latitude), ]
# Ensure data is not NULL and has the required columns
#if (!is.null(data) && all(c("fire_location_longitude", "fire_location_latitude", "fire_spread_rate", "

  # Print debugging information

  # Filter out rows with missing lat/long
  #data <- data[!is.na(data$fire_location_longitude) & !is.na(data$fire_location_latitude), ]

  #leaflet(data) %>%
    #addTiles() %>%
    #addCircleMarkers(~fire_location_longitude, ~fire_location_latitude,
                     #radius = ~fire_spread_rate * 0.1,
                     #color = ~ifelse(SpreadCategory == "High", "red", "green"),
                     #fillOpacity = 0.5,
                     #popup = ~paste("Spread Rate:", fire_spread_rate, "<br>",
                                    #"Temperature:", temperature)) %>%
    #setView(lng = mean(data$fire_location_longitude, na.rm = TRUE),
            #lat = mean(data$fire_location_latitude, na.rm = TRUE),
            #zoom = 6) %>%
    #addLegend("bottomright",
              #colors = c("red", "green"),
              #labels = c("High Spread Rate", "Low Spread Rate"),
              #title = "Spread Rate Category")
#} else {
  #print("Data is NULL or required columns are missing.")
#}
```

Added clustering Points for Better Performance on this geospatial visualization. This will interactively shows the area where the fire spread rate and temperature is high or low.

Hypothesis testing:

```r
# Extract the p-value for temperature in the transformed model
p_value_log <- summary(log_model)$coefficients[2, 4]

# Hypothesis testing
if (p_value_log < 0.05) {
  print("Reject the null hypothesis. There is a statistically significant relationship between temperat
} else {
  print("Fail to reject the null hypothesis. There is no statistically significant relationship between
}
```

```
## [1] "Reject the null hypothesis. There is a statistically significant relationship between temperatu
```

```r
# Print p-value for confirmation
print(paste("P-value:", p_value_log))
```

```
## [1] "P-value: 1.1251791730511e-317"
```

There is a statistically significant relationship between temperature and fire spread rate (log-transformed).

The analysis of the residual plot and Q-Q residual plot suggests that the linear regression model is not an appropriate fit for the data. The residual plot indicates a poor linear relationship, and the Q-Q residual plot reveals that the data is not normally distributed, with a right-skew and the presence of outliers. Given these observations, alternative modeling approaches, such as non-linear regression or robust regression techniques, should be considered to better capture the underlying patterns in the data and avoid the reliance on normality assumptions.