Analyzing Relationship Between Reddit Sentiment & Weather

Problem Statement

The problem we are looking to address and analyze is to see whether or not there is a relationship between weather and the sentiment of reddit comments in a subreddit. In order to analyze this problem, we broke down and refined the idea to these four main questions: (1) Is subreddit post sentiment impacted by rain?, (2) Is subreddit post sentiment impacted by snow?, (3) Is subreddit post sentiment impacted when days are hotter? , and (4) Is subreddit post sentiment impacted when days are cold?

In order to address these questions, we further broke down these questions to be able to perform various tests to answer and conclude on them. We discuss these questions further in the "Results and Findings" section.

Data Gathering, Cleaning, and Preparation

Subreddit Data

We decided to look specifically at city subreddits since we can determine a particular city's weather at the time of users' comments. We decided to limit the subreddits to Canada and the US and wanted to look at cities with more subscribers, since that would provide more comments to look at. The subreddit data was gathered from the SFU compute cluster, which provided all reddit comments from years 2015 to 2021. We used Spark to extract the comments for the relevant city subreddits we wanted to look at and split the data into separate files based on city. In terms of cleaning the subreddit data, we removed all removed/deleted comments from the data.

Weather Data

The weather data we used was gathered from GHCN. From reviewing the GHCN data, we found a lot of weather station data collected from the year 1750 until present. We have decided to work on the data from 2015 to 2021 to align with the subreddit data. From there, we have used the GHCN website to view each city we selected and chose a nearby weather station for each city that had the most coverage of data for those years.

We selected the following cities based on subreddit size and weather station coverage: New York City, Los Angeles, Boston, Chicago, Seattle, Atlanta, San Francisco, Toronto, Vancouver, Calgary, Montreal.

For preparing the data, GHCN noted that the precipitation measurements were recorded as tenths of mm, and the temperatures were recorded as tenths of Celsius. Therefore, we divided the precipitation and temperature values by 10 to get the correct values in mm and Celsius, respectively. In addition, we removed temperature data with all 9's in the field, such as 9999, as GHCN indicated this meant there was no measurement. To analyze the data, we took the average of the minimum and maximum temperatures, and labeled it as cold weather if the average temperature is less than or equal to 13 degree Celsius, otherwise, we labeled it as hot weather.

Sentiment Analysis

To analyze reddit comment sentiment, we used an external python library called VADER Sentiment Analysis which is an open source lexicon and rule-based sentiment analysis tool for social media posts. The library analyzes the text and provides a compound score which is normalized between -1 (most negative) and 1 (most positive). Based on the library's documentation, the scores can be categorized into thresholds as follows:

- Compound score >= 0.05 is positive sentiment
- Compound score > -0.05 and compound score < 0.05 is neutral sentiment
- Compound score <= -0.05 is negative sentiment

Results and Findings

In order to answer our problem statement, we broke down the problem into various questions where we performed statistical analysis to see results and try to come to a conclusion based on the findings. We discuss the questions we address and the type of analysis we have done in order to see results below.

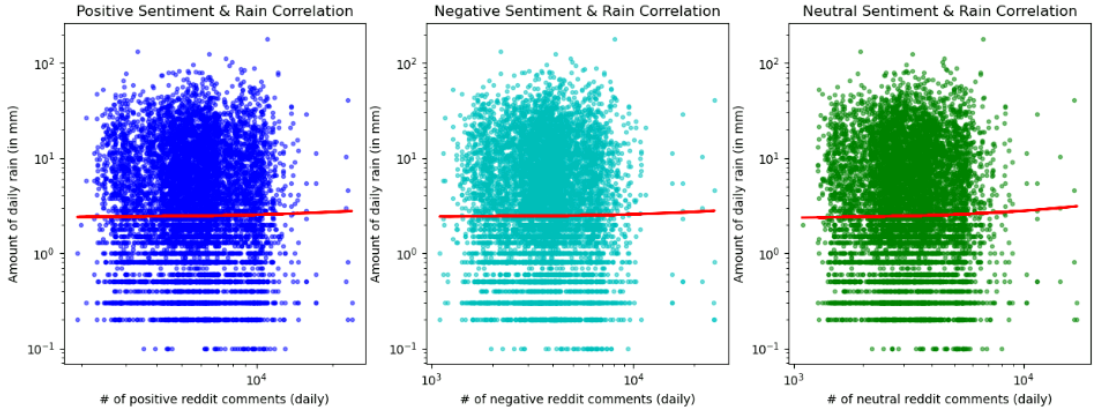
Is post sentiment impacted by rain or snow?

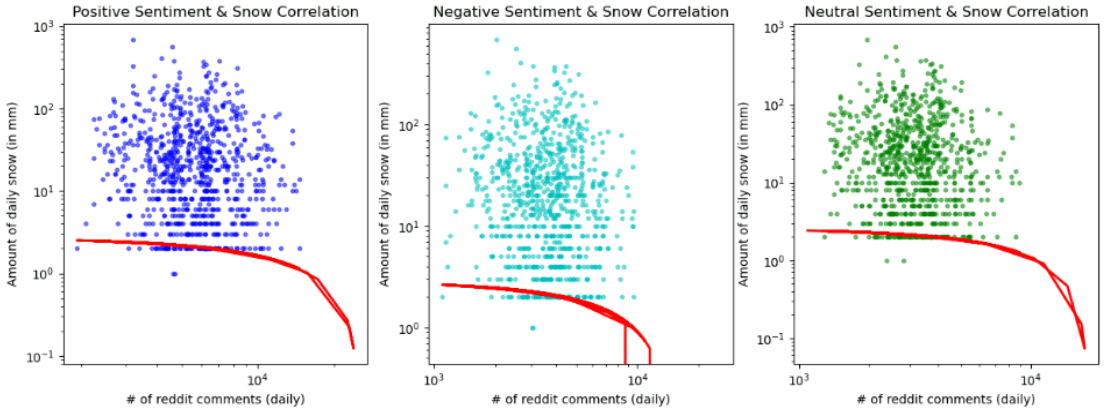
Is there a linear relationship between sentiment of posts and amount of rain/amount of snow?

To analyze if there was any linear relationship between sentiment of posts and amount of rain/snow, we performed linear regression on the number of posts per sentiment against the amount of rain per day (in mm) and amount of snow per day (in mm). The results from the testing were as follows:

	Rain			Snow		
Sentiment	Positive	Negative	Neutral	Positive	Negative	Neutral
P-Value	0.38774	0.54387	0.17277	0.06106	0.00971	0.15237
R-Value	0.00554	0.00389	0.00875	-0.01502	-0.02073	-0.01147

Looking at all three sentiments for both rain and snow, we can see that the p-values are all higher than our alpha of 0.05, with the exception of negative posts for snow. For this case, we could potentially reject the null hypothesis, but looking at the graph as well, there does not seem to be any indication of a linear relationship. For all other cases excluding this exception, we are not able to conclude that the regression slope between the sentiment and the amount of rain is not 0. In addition, for both rain and snow, we can see for all three sentiments, they all have r-values that are very close to 0. This indicates to us that there is little or no linear relationship between the amount of rain or snow and # of posts by sentiment. We plot the data points and best fit-line onto the graphs below, and log-scaled it to better show the data distribution. The graphs further show us that there does not seem to be any linear relationship between the amount of rain or snow and sentiment.





Is there an association between subreddit post sentiment and having rain/having snow?

Another method we wanted to test was to look at the data categorically. We used the Chi-Square test to analyze if having rain or snow had any effect on the type of post sentiment. For the contingency table, we used the following categories: for sentiments we used, positive, negative and neutral and for rain and snow data, we categorized the data as either has rain/has snow (> 0 mm) or no rain/no snow (0 mm). The p-value of the

Chi-Square test was $7.3859 \times e^{-17}$ for rain and was $1.443 \times e^{-271}$ for snow. Since both the p-values are below the alpha of 0.05, we can reject the null hypothesis (it doesn't matter if there is rain/snow or not, it has no impact on sentiment), and this indicates that having rain/snow or no rain/snow does potentially have some effect on what the reddit comment sentiment is. These results will need to be analyzed together with our other testing to understand these results further.

Is there a different number of positive, negative, and neutral comments when it is raining/snowing?
We also performed ANOVA tests to see if the average number of reddit comments for each sentiment on rainy/snowy days is different. From there, we performed post-hoc analysis with Tukey's HSD test to get a better idea of which groups had differences, if any. To specifically look at days with rain, we separated the data to only include days that had rain > 0 mm and similarly, for snow, we only included days with snow measurements > 0. From the ANOVA tests, the p-values for rainy days or snowing days were both essentially 0. This result indicates that there is a difference in means between the sentiments. The Tukey's HSD test provided this information:

Rain				Snow			
Group 1	Group 2	Mean Diff.	Reject	Group 1	Group 2	Mean Diff.	Reject
Negative	Neutral	-619.9164	True	Negative	Neutral	-506.0542	True
Negative	Positive	2039.8803	True	Negative	Positive	2083.7762	True
Neutral	Positive	2659.7968	True	Neutral	Positive	2589.8303	True

From these results, we can see when it is snowing or raining, there are more positive sentiment comments on average. However, in order to see if this result is significant, we wanted to compare these results to data where there is no rain or snow as well to see if this difference is truly being driven by the rain/snow. We run the ANOVA test below to understand this result better.

Is there a different number of positive, negative, and neutral comments?
From the previous ANOVA test's results, we wanted to verify if this result of higher mean of positive posts was specific to only rain/snow data or if this is a trend of the subreddit data overall. From running the ANOVA tests containing data with both rain days and no rain days and snow and no snow days, the p-values are essentially 0, which indicates that the null hypothesis cannot be true and that there is a difference in means. Therefore, we performed a Tukey's HSD test as well. The main significant results from these tests are the following:

Rain				Snow			
Group 1	Group 2	Mean Diff.	Reject	Group 1	Group 2	Mean Diff.	Reject
No Rain - Negative	No Rain - Neutral	-637.98	True	No Snow - Negative	No Snow - Neutral	-628.12	True
No Rain - Negative	No Rain - Positive	2015.57	True	No Snow - Negative	No Snow - Positive	1981.23	True
No Rain - Neutral	No Rain - Positive	2653.55	True	No Snow - Neutral	No Snow - Positive	2609.35	True
No Rain - Negative	Rain - Negative	20.16	False	No Snow - Negative	Snow - Negative	-97.26	False
No Rain - Neutral	Rain - Neutral	38.23	False	No Snow - Neutral	Snow - Neutral	24.80	False
No Rain - Positive	Rain - Positive	44.47	False	No Snow - Positive	Snow - Positive	5.28	False

From these results, we can see that on days that have no rain, positive posts also have a higher mean compared to negative and neutral posts. When comparing the no rain days and rain days, we are not able to conclude that

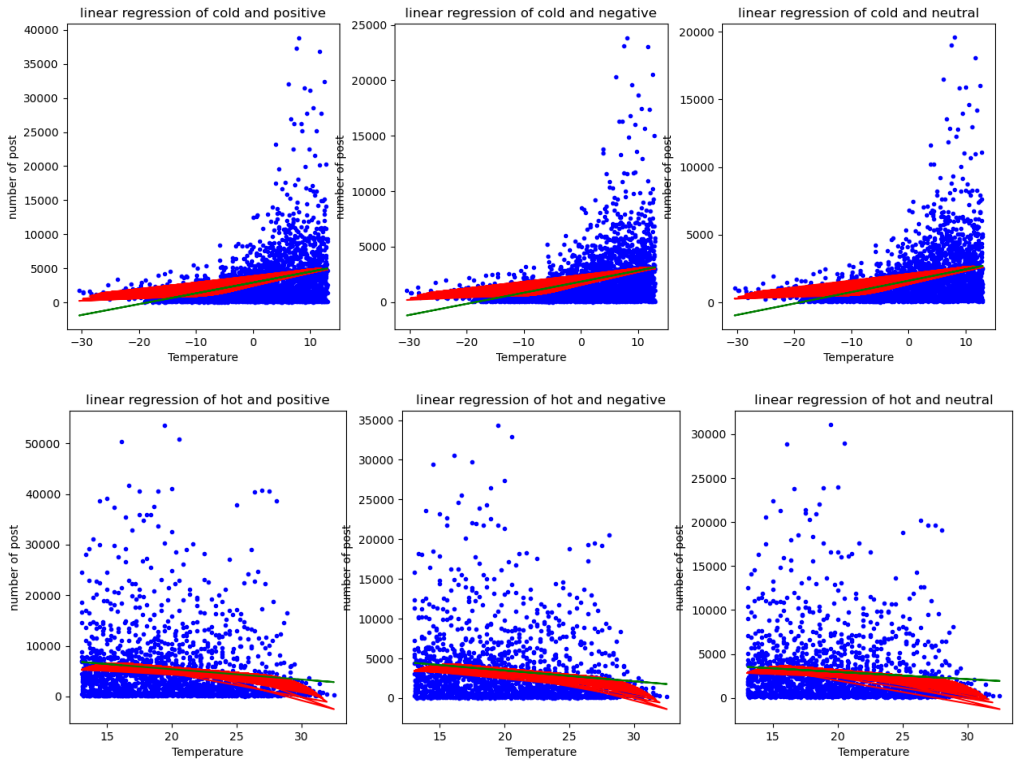
there is a difference in means for each sentiment. (i.e. we are not able to conclude that the rainy day means are higher than the non-rainy days). We see similar results when looking at the snow and no snow data as well. Therefore, this indicates to us that the subreddits have more positive posts in general, and the higher amount of positive posts vs negative/neutral posts is not likely due to the rain or snow.

Is post sentiment impacted by cold or hot weather?

Is there a relationship between positive/neutral/negative posts and temperature in cold and hot weather?

In order to analyze if there is any relationship between post sentiment and cold or hot weather, we performed linear and polynomial regression on the number of posts per sentiment against the temperature in cold weather per city. The results were as follow:

	Cold			Hot		
	Positive	Negative	Neutral	Positive	Negative	Neutral
p-value	$4.04 \times e^{-71}$	$1.05 \times e^{-71}$	$1.37 \times e^{-73}$	$6.47 \times e^{-8}$	$9.10 \times e^{-9}$	$6.54 \times e^{-8}$
r-value	0.352765	0.354158	0.358586	-0.134251	-0.142644	-0.134202



In the graphs, the blue dots are the number of posts for each sentiment, and the red line is the prediction from a polynomial regression result, the green line is the prediction from a linear regression model. From the graphs, we can see that the polynomial regressions generated some predictions that are bouncing around, which shows that the data is unpredictable. In addition, the p-values of the linear regressions are extremely small for both hot and cold temperatures. Therefore, we are not able to reject the null hypothesis that there are no relationships between either positive, neutral, or negative posts with cold/hot weather.

Is there an association between post sentiment and different weather?

To analyze if there is any association between post sentiment and different weather categorically, we have performed a Chi-Square test on the number of posts per sentiment against hot and cold categories. After performing the test, we have got a p-value of $2.379504 \times e^{-188}$. We can see that they both got an extremely low p-value, which means that we can not reject the null hypothesis, which indicates that there is no difference if it is in the category of hot or cold temperature, this has no impact on sentiment of positive, negative, and neutral posts.

Is there a different number of positive, negative, and neutral comments in cold weather?
To analyze if there is a different number of positive, negative, and neutral comments in cold and hot weather, we have performed an ANOVA test, which resulted in a p-value of $5.006185 \times e^{-192}$. Therefore, we performed a Tukey's HSD test. The results were as follow:

		Cold		Hot	
Group 1	Group 2	Mean Diff.	Reject	Mean Diff.	Reject
negative	neutral	-327.2338	True	-335.0069	True
negative	positive	1434.9062	True	1233.9266	True
neutral	positive	1762.14	True	1568.9335	True

From the table above, we could generally reject the null hypothesis that claims there is no difference in means between positive, negative, and neutral posts for both cold and hot weather.

Do individual cities have relationships between weather and post sentiment?
Another consideration we had was that we wanted to see if individual cities had potentially different relationships with post sentiment and weather. For example, if rain affected post sentiment more in Vancouver versus New York. We ran similar tests as above such as linear regression, ANOVA tests, and Chi-Square tests for the different weather types and temperatures for Vancouver and New York individually. The results we found were very similar to the results above and we didn't note any significant results for those two cities.

Machine Learning Predictions
In addition to statistical analysis we did, we have also performed a machine learning random forest model to see the prediction result. The idea is that if the model was able to accurately predict the post sentiment based on the weather features we provided, this would indicate to us that there must be some relationship that exists. For the model, we have set the city and weather data (PRCP, SNOW, TMAX, TMIN, TAVG) as the features, and the sentiment result as the label. However, after testing the accuracy on the validation set, we have got a pretty low score on it of 45.3411%. This low accuracy score indicates to us that there does not appear to be a strong relationship since the model is not able to predict the label based on the features we provided.

Conclusion
Based on the results and findings, we were not able to find any significant results that indicated that the weather had an impact on reddit comment sentiment. We did find in the Chi-Square test of rain and snow weather, that the test did indicate the rain/snow may impact the type of post sentiment there would be. However, from looking at the combined results from our other statistical tests like ANOVA and linear regression, the majority of the data is telling us that this is not the case. Overall, from the statistical analysis we did, the tests indicate that there does not seem to be a relationship between weather and comment sentiment. This was further confirmed by looking at cities individually, and also by testing using machine learning prediction as well.

One thing we did note from our tests is that there does seem to be a trend that there are more positive comments in the subreddits compared to neutral and negative posts. This result is likely due to the fact that reddit subreddits are heavily moderated, so a lot of the negative posts likely have already been deleted or removed.

Limitations
- One of the data points we wanted to analyze from GHCN was also to look at sunny days. GHCN has some measurements for the amount of sun in a day, but upon reviewing the data, we noted that a lot of the weather stations did not have this measurement available or recorded.
- If we had more time, we also would have liked to test more cities individually to see if there were any significant results in terms of relationships of post sentiment and weather for specific cities.

Project Experience Summary

Leslie:

- Conducted a data science project to analyze the potential relationship between weather patterns and Reddit comment sentiment.
- Performed extract-transform-load tasks to gather and transform subreddit comment data using Spark to prepare data for analysis
- Conducted data cleaning tasks on weather and subreddit data to ensure data is valid and usable for project analysis
- Implemented vaderSentiment library to perform sentiment analysis on Reddit comments.
- Utilized Spark and Pandas for efficient data processing and manipulation of large Reddit comment datasets.
- Performed statistical analysis with statsmodels to explore relationship between weather and reddit comment sentiment and to identify potential trends and patterns.

Angus:

- Performed data science technique to extract, transform, and load datasets.
- Utilized different machine learning techniques to make a better prediction on the sentiment by the weather features.
- Performed different searching techniques to build better datasets.
- Conducted different statistical test to explore relationship between datasets.