Bay Area Bike Rental Operation Research

Data Cleaning & Analysis

# Introduction

This data analysis serves as the first step in developing a predictive model that will predict the number of bikes leaving each station and the number of bikes being returned to each station in the next three days. It is split up into two parts, one which focusses on 97.5% of the data, excluding the very longest trips such that the maximum duration is 85 minutes, while the other includes these extreme values to reflect the fact that individuals can subscribe to possess a single bike for months at a time.

# Data Cleaning & Remove Extreme Values

There are three datasets to be examined and cleaned. One on Bay Area stations, one on bike trips over the course on one year, and one on weather data for five cities over the course of one year.

Most variables had missing values coded appropriately or were easily modified so that empty values were recorded with an NA value. The only categorical variable with a significant amount of missingness was Weather Events. Variables that were incorrectly coded as character data types were converted to POSIX or numeric values.

### Station Data

All 70 stations had their associated city recorded. There are no missing cities. Not surprisingly the most frequent city cited in the dataset is the metropolitan San Francisco.



Figure 1. City frequency within the Station dataset

The station data also lists their latitude, longitude, and dock counts. The latitudes are grouped into 2 clusters while the longitudes are grouped into 3 clusters, as can be seen by the distribution of stations in Figure 39.



Figure 2. Frequency of different latitude, longitude, and station dock count values

### Trip Data

When trip data was examined, all 70 stations were represented. The most frequent station from which trips started and ended was San Francisco Caltrain (Townsend at 4th).

A chart of a number of cities

Description automatically generated with medium confidence

Figure 3. Frequency of station names that start bike trips within the Station dataset

A chart of a train

Description automatically generated with medium confidence

Figure 4. Frequency of station names that end bike trips within the Station dataset



Figure 5. Frequency of bike trips taken by either a subscriber or a customer

The majority of bike users are subscribers rather than customers. This may indicate differences in resident versus visitor/tourist populations, though such speculative detail is absent in the data. All users are either customers or subscribers.

### Weather Data



Figure 6. Frequency of weather events within the Weather dataset

As stated previously, 1473 out of the 1825 weather events were not recorded, and it was unclear whether the weather was merely sunny, or events such as rain or fog had not been recorded. There is some evidence to the veracity of the Rain reporting, as it rained on 52 days in 2014 in the Bay Area 1.

To account for this, weather events were imputed using the mean visibility to determine fog, haze, and mist, and the precipitation to impute rainfall.

The criteria for imputation were as follows 2:

Fog: visibility less than 1 km (0.62 miles)

Mist: visibility between 1 km (0.62 miles) and 2 km (1.2 mi)

Haze from 2 km (1.2 miles) to 5 km (3.1 miles)

Temperatures at 14 F (-10°C) is too cold for the air to contain super-cooled moisture, and no fog, haze, or mist is possible. No temperatures recorded reached this level.

Rain: any amount of precipitation, even trace

Thunderstorm: hot temperatures, cloud cover, wind-speed, but highly variable, and thus could not be imputed.

50 events were able to be imputed, though this still left 1423 events out of 1825 as “not recorded”. Notably, the imputation split up Fog-Rain into more precise categories of mist, haze, and fog. Furthermore, because thunderstorms could not be imputed from other recorded phenomena (temperature, precipitation, etc.), it is unclear how many of the new Rain, Haze-Rain, Mist-Rain, or Fog-Rain are incorrectly imputed. The one instance of Rain-Thunderstorm is not imputed but is pulled over from the recorded events. Thus, these imputed events were not used within the analysis.



Figure 7. Frequency of imputed weather events within the Weather dataset



Figure 8. Frequency of different weather variable values



Figure 9. Frequency of cities with recorded weather data

Every day of each of the 5 cities was accounted for weather-wise (though again, it is unclear how reliable the reporting of the weather is).

Regardless of city, the weather variables were plotted for a single year, and reveal expected seasonal patterns.

Temperature rises over the course of summer before falling near winter (Fig 10-12).



Figure 10. Maximum temperature readings for 5 different Bay Area cities over an entire year



Figure 11. Mean temperature readings for 5 different Bay Area cities over an entire year



Figure 12. Minimum temperature readings for 5 different Bay Area cities over an entire year

The Bay Area primarily has a visibility of 10 miles, even across the maximum, mean, and minimum plots (Figure13-15). It is unclear how visibility was calculated, though it is likely based on the rate that laser light scatters in water vapour 3. It is also uncertain when the measurements were made.



Figure 13. Maximum visibility readings for 5 different Bay Area cities over an entire year



Figure 14. Mean visibility readings for 5 different Bay Area cities over an entire year



Figure 15. Minimum visibility readings for 5 different Bay Area cities over an entire year

With some exceptions, most wind speeds in the Bay Area are below 30 mph. However, the extreme values seen in Figure 16 are within the range of wind speeds seen in tropical storms, as are the gusts in Figure 18. Tropical storms can have wind speeds in excess of 50 mph and gust speeds in excess of 65 mph 4. These days were removed from the analysis.



Figure 16. Maximum wind speed values for 5 different Bay Area cities over an entire year



Figure 17. Mean wind speed values for 5 different Bay Area cities over an entire year



Figure 18. Maximum gust speed readings for 5 different Bay Area cities over an entire year



Figure 19. Precipitation measurements for 5 different Bay Area cities over an entire year



Figure 20. Cloud cover readings for 5 different Bay Area cities over an entire year

# Analysis Sans Extreme Values

### Outlier and Extreme Value Removal

|  |  |
| --- | --- |
| A white text box with black text  Description automatically generated | A graph with a line and dots  Description automatically generated with medium confidence |
| A diagram of a bike trip duration  Description automatically generated | A diagram of a bike trip duration  Description automatically generated |

Figure 21. Sequential cleaning of data, excluding extreme values

The vast majority of trips were less than 10 minutes long, as indicated by the tiny interquartile range shown in the box plots.

Outliers and extreme values were removed from the dataset. Outliers were considered bikes that have trip durations longer than a month but were not taken out by subscribers. These may be bikes that were forgotten to be returned. Only one bike was an outlier under this metric and was taken out by a regular customer for 287839 minutes (~200 days long), while the next longest duration was only 12007 minutes (~8 days).

From there, extreme values were removed. These are values that while technically feasible, were not considered for this analysis. The upper 2.5% of trips were removed, leaving the longest trip to now be 85 minutes long. Finally, cancelled trips were considered to be any trips that lasted less than 3 minutes, and so were excluded leaving the cleaned dataset containing 312,062 trips.

Notably, the Duration variable was considered unreliable, as it recorded trips lasting longer than the start and end times would suggest. Therefore, a new variable was constructed using the difference in time between the start and end time-dates.



Figure 22. Frequency of subscribers and customers for excluded trips (duration > 85 minutes)

Of these removed trips, the vast majority were from customers, rather than subscribers. This was unlike the majority of the data. It is unclear what conclusions can be drawn from this. Though this does indicate that while possible that a customer can check out a bike for up to a month, it is interesting that the group of trips longer than 85 minutes was comprised primarily of customers rather than having just as many subscribers who could check out the bike for months at a time. One possibility is that if the speculation that many of the customers are tourists rather than residents is true, then this may reflect the fact that they might not own a car.

### Rush Hour Analysis

To determine which bikes and stations should be the focus of maintenance, it is necessary to understand which are used the most. One key component is determining “rush” hours, and identifying which bikes and stations are under the maximum strain of a transit system.



Time of Day

Figure 23. Number of trips per the same 15-minute interval across 1 year (rush hours in red)

Rush hours are defined in this model as the busiest 20% of the day, the quarter of time when the most bikes are in concurrent use. The cutoff, the 80th percentile of trips per 15-minute interval, is 10380 trips over a single year.

In the Bay Area, there are two Rush Hour periods, one in the morning from 7:00-9:15 EST and one in the afternoon from 15:45-18:15 EST. This is the time when the most bikes are being used, as well as which stations are being used during peak times. Rush-hour stations should be a priority for maintenance due to their importance, but likewise, should be repaired/maintained during off-hours. The top 10 starting stations and ending stations are listed in Table 1 and Table 2.

San Francisco Caltrain (Townsend at 4th) being a major stop is not surprising given it is a major terminal for the San Francisco Caltrain commuter train and MUNI transit, for a total of 17 different tracks 5. It dominates the top 10 list of starting and ending stations.

Table 1: Top 10 starting stations for rush hour

|  |  |
| --- | --- |
| Starting Station | Number of Trips Per Year |
| San Francisco Caltrain (Townsend at 4th) | 2126 |
| San Francisco Caltrain 2 (330 Townsend) | 715 |
| Temporary Transbay Terminal (Howard at Beale) | 635 |
| Grant Avenue at Columbus Avenue | 528 |
| Harry Bridges Plaza (Ferry Building) | 520 |
| San Jose Diridon Caltrain Station | 435 |
| Steuart at Market | 406 |
| Market at 10th | 342 |
| Townsend at 7th | 243 |
| South Van Ness at Market | 240 |

Table 2: top 10 ending stations for rush hour

|  |  |
| --- | --- |
| Ending Station | Number of Trips Per Year |
| San Francisco Caltrain (Townsend at 4th) | 1067 |
| 2nd at Townsend | 564 |
| Market at Sansome | 550 |
| Townsend at 7th | 494 |
| Temporary Transbay Terminal (Howard at Beale) | 466 |
| Embarcadero at Folsom | 448 |
| San Francisco Caltrain 2 (330 Townsend) | 431 |
| San Jose Diridon Caltrain Station | 352 |
| Mountain View Caltrain Station | 349 |
| Steuart at Market | 348 |

Interestingly enough, the most common starting and ending station on the weekends is Embarcadero at Sansome. This station is right by Fisherman’s Wharf, a major tourist and recreation destination 6. Likewise, the Harry Bridges Plaza (Ferry Building) station is in the middle of a plaza frequented by many artists, skaters, and tourists 7.

Table 3: top 10 starting stations for weekend

|  |  |
| --- | --- |
| Starting Station | Number of Trips Per Year |
| Embarcadero at Sansome | 2868 |
| Harry Bridges Plaza (Ferry Building) | 2741 |
| Embarcadero at Bryant | 1488 |
| Market at 4th | 1477 |
| 2nd at Townsend | 1459 |
| Powell Street BART | 1341 |
| San Francisco Caltrain (Townsend at 4th) | 1296 |
| Grant Avenue at Columbus Avenue | 1191 |
| Market at 10th | 994 |
| Market at Sansome | 988 |

Table 4: Top 10 ending stations for weekend

|  |  |
| --- | --- |
| Ending Station | Number of Trips Per Year |
| Embarcadero at Sansome | 2983 |
| Harry Bridges Plaza (Ferry Building) | 2852 |
| Market at 4th | 1690 |
| San Francisco Caltrain (Townsend at 4th) | 1581 |
| Powell Street BART | 1540 |
| 2nd at Townsend | 1520 |
| Embarcadero at Bryant | 1268 |
| Steuart at Market | 1116 |
| Townsend at 7th | 1026 |
| Civic Center BART (7th at Market) | 991 |

Further exploration can highlight who is using the bike system during rush hours and the weekends, whether it is people going to work, or going out for leisure. By knowing what stations are being used for, it can determine which should be a priority. For example, San Francisco Caltrain (Townsend at 4th) is in the top 10 list of starting and ending stations for both rush hours and weekends, and by focussing on repairing it, both workers and leisure-goers can be advantaged.

### Overall & Average Bike Usage

When it comes to overall bike usage, when stratified by month, approximately 70% of the time there are some bikes in use. This is approximately constant across all 12 months (Figure 24.). Notably, this is a measure that includes at least a single bike on the road, and thus is not the most reliable measure.



Figure 24. Percent of minutes per month when at least a single bike is on a trip

However, individual bikes are used, on average, approximately 1% of a month’s duration, or about 7.2 hours (Figure 4.). Bike usage is highest during the summer months. This follows the warmer weather of the summer months (Figure 10-12).



Figure 25. Average bike usage per month for individual bikes

Per month, the top 20 bikes used vary. Depending on the month that repairs are occurring, it may be reasonable to focus on the bikes that are going to be used in the upcoming months, that way they are in top shape for customers and subscribers. Likewise, knowing when each bike is in use can help for scheduling repairs so that preferential bikes are not unavailable. The top bikes are also likely to be in greater need for repair.

Table 5: Top 20 bikes (IDs listed) used per month

| Rank | Jan. | Feb. | Mar. | Apr. | May | Jun. | Jul. | Aug. | Sept. | Oct. | Nov. | Dec. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 596 | 357 | 556 | 395 | 528 | 349 | 467 | 516 | 300 | 459 | 304 | 576 |
| 2 | 473 | 475 | 335 | 386 | 339 | 298 | 511 | 282 | 318 | 590 | 302 | 306 |
| 3 | 335 | 253 | 289 | 586 | 488 | 388 | 497 | 496 | 266 | 57 | 363 | 397 |
| 4 | 337 | 466 | 583 | 333 | 410 | 170 | 321 | 395 | 536 | 510 | 573 | 434 |
| 5 | 436 | 465 | 586 | 588 | 537 | 352 | 599 | 318 | 558 | 259 | 576 | 555 |
| 6 | 506 | 517 | 303 | 360 | 299 | 467 | 550 | 285 | 555 | 578 | 353 | 486 |
| 7 | 561 | 580 | 547 | 560 | 444 | 686 | 388 | 343 | 361 | 242 | 511 | 379 |
| 8 | 375 | 370 | 497 | 256 | 276 | 304 | 309 | 440 | 392 | 382 | 372 | 288 |
| 9 | 573 | 266 | 418 | 464 | 249 | 432 | 201 | 170 | 537 | 593 | 402 | 409 |
| 10 | 313 | 249 | 598 | 305 | 534 | 437 | 567 | 560 | 324 | 379 | 561 | 517 |
| 11 | 341 | 412 | 414 | 345 | 403 | 282 | 263 | 435 | 57 | 298 | 329 | 420 |
| 12 | 585 | 255 | 471 | 442 | 544 | 548 | 455 | 121 | 362 | 215 | 435 | 289 |
| 13 | 523 | 261 | 316 | 592 | 519 | 276 | 480 | 417 | 528 | 334 | 194 | 488 |
| 14 | 587 | 392 | 589 | 424 | 447 | 378 | 531 | 379 | 459 | 124 | 392 | 394 |
| 15 | 352 | 242 | 266 | 591 | 579 | 583 | 477 | 599 | 21 | 383 | 244 | 389 |
| 16 | 451 | 514 | 459 | 417 | 124 | 309 | 603 | 627 | 174 | 254 | 351 | 496 |
| 17 | 342 | 472 | 596 | 532 | 645 | 480 | 598 | 589 | 257 | 355 | 411 | 376 |
| 18 | 343 | 301 | 585 | 387 | 385 | 243 | 528 | 495 | 194 | 687 | 440 | 396 |
| 19 | 374 | 415 | 536 | 452 | 529 | 492 | 509 | 378 | 389 | 263 | 145 | 467 |
| 20 | 390 | 315 | 387 | 462 | 328 | 571 | 604 | 536 | 519 | 567 | 328 | 573 |

### Trip & Weather Correlation

When it came to examining the relationship between weather and trip information, focus was given to trip duration, city, and the number of trips. The number of trips was calculated by examining how many strips started under the same weather conditions.

These values were chosen because it can aid in determining where repairs should be focussed. If weather correlates with how long a bike is used and/or how often bikes are used, then that will impact the repair schedule of a bike. Furthermore, if certain weather conditions correlate more with one city than another, and that condition is found to correlate with bike usage, then that particular city may be an area of focus for repair efforts.

As can be seen in Figure 26, trip duration was found to be largely uncorrelated with the weather predictors. The highest magnitude correlation for trip duration was with maximum visibility, which if it were more significant might give room for hypotheses on how recreational use of a bike may be extended when visibility is higher due to ameliorated safety concerns, but ultimately the correlation coefficient is only 0.04.

The number of trips that occur given the weather conditions have stronger correlations. Temperature has the highest correlation, indicating that warmer weathers may inspire more trips, which would support the findings in Figure 24. Precipitation correlates negatively with the number of trips. This follows intuition, as rain can be a dangerous weather condition for bike riding. Notably, these correlations are not particularly strong, indicating the limited efficacy of this particular data analysis.



Figure 26. Correlation coefficients for weather and trip variables for all users



Figure 27. Correlation coefficients for weather and city variables for all users

Once more, the weather does not correlate strongly with any particular city in the Bay Area. The city of Palo Alto has a substantial correlation with maximum visibility in miles, though given the lack of any other substantial correlations, including the other visibility terms, it is not unfair to consider this to be likely a coincidence.

Given that trips are divided between customers and subscribers, correlations were performed on each group to attempt to discern any behavioural differences between the two groups. Results were very similar to the aggregated scenario. The number of trips had stronger positive and negative correlations, but not by much.



Figure 28. Correlation coefficients for weather and trip variables for all customers



Figure 29. Correlation coefficients for weather and trip variables for all subscribers

The trip-weather correlations for subscribers followed the aggregated scenario, which is not surprising given the overwhelming number of bike users are subscribers.

Despite this, it is clear that customers and subscribers behave differently. As can be seen in Table 6, the ratio between customers and subscribers is not constant between weekends. While the number of customers is on the same order of magnitude between weekends and weekdays, the same is not true for subscribers, adding evidence that subscribers are more likely to be workers than customers.

Table 6: Difference is customer and subscriber trips across weekdays and weekends

|  |  |  |  |
| --- | --- | --- | --- |
|  | Customer Trips | Subscriber Trips | Customer/Subscriber |
| Weekday | 24,733 | 256,697 | 0.096 |
| Weekend | 16,295 | 20,393 | 0.799 |

# Expanded Analysis

### Outlier Removal

The only difference between the two analyses was that in the expanded analysis, all possible values were considered. It is conceivable that a customer could rent a bike for up to 8 days, and so the only outlier was a trip duration longer than a month that belonged to a customer rather than a subscriber. Like the first analysis, cancelled trips were considered to be any trips that lasted less than 3 minutes, and so were excluded leaving the cleaned dataset containing 320,283 trips

|  |  |
| --- | --- |
| A white text box with black text  Description automatically generated | A line graph with black dots and white text  Description automatically generated with medium confidence |

Figure 30. Sequential cleaning of data, including extreme values

### Rush Hour Analysis



Figure 31. Number of trips per the same 15-minute interval across 1 year (rush hours in red)

Rush hours are defined in this model as the busiest 20% of the day, the quarter of time when the most bikes are in concurrent use. The cutoff, the 80th percentile of trips per 15-minute interval, is 12540 trips over a single year.

In the Bay Area, there are two Rush Hour periods, one in the morning from 7:15-9:00 EST and one in the afternoon from 15:00-18:00 EST. These rush hour periods are very similar to the ones produced in the restricted analysis. This is the time when the most bikes are being used, as well as which stations are being used during peak times. Rush-hour stations should be a priority for maintenance due to their importance, but likewise, should be repaired/maintained during off-hours. The top 10 starting stations and ending stations are listed in Table 7 and Table 8.

Once again, the San Francisco Caltrain (Townsend at 4th) station is the predominant station during rush hours, however now the Harry Bridges Plaza (Ferry Building) station is the second most common starting station during rush hours, though why this may be is unclear.

Table 7: Top 10 starting stations for rush hour

|  |  |
| --- | --- |
| Starting Station | Number of Trips Per Year |
| San Francisco Caltrain (Townsend at 4th) | 2460 |
| Harry Bridges Plaza (Ferry Building) | 1145 |
| San Francisco Caltrain 2 (330 Townsend) | 1097 |
| Temporary Transbay Terminal (Howard at Beale) | 1000 |
| Grant Avenue at Columbus Avenue | 595 |
| Steuart at Market | 559 |
| 2nd at Townsend | 449 |
| Market at 10th | 424 |
| South Van Ness at Market | 319 |
| Civic Center BART (7th at Market) | 312 |

Table 8: top 10 ending stations for rush hour

|  |  |
| --- | --- |
| Ending Station | Number of Trips Per Year |
| San Francisco Caltrain (Townsend at 4th) | 843 |
| Market at Sansome | 733 |
| 2nd at Townsend | 660 |
| Townsend at 7th | 644 |
| Embarcadero at Sansome | 551 |
| Temporary Transbay Terminal (Howard at Beale) | 497 |
| Embarcadero at Folsom | 468 |
| Clay at Battery | 456 |
| Steuart at Market | 422 |
| Beale at Market | 404 |

The top 10 starting and ending stations for the weekend are almost identical to those in the restricted case. As the majority of additional trips in the expanded case belong to customers rather than subscribers, it may support the notion that they are more likely to pursue leisure than subscribers, being tourists instead of workers.

Table 9: top 10 starting stations for weekend

|  |  |
| --- | --- |
| Starting Station | Number of Trips Per Year |
| Harry Bridges Plaza (Ferry Building) | 3154 |
| Embarcadero at Sansome | 3118 |
| Market at 4th | 1656 |
| Embarcadero at Bryant | 1568 |
| 2nd at Townsend | 1537 |
| Powell Street BART | 1473 |
| San Francisco Caltrain (Townsend at 4th) | 1359 |
| Grant Avenue at Columbus Avenue | 1296 |
| Powell at Post (Union Square) | 1088 |
| Market at Sansome | 1075 |

Table 10: Top 10 ending stations for weekend

|  |  |
| --- | --- |
| Ending Station | Number of Trips Per Year |
| Embarcadero at Sansome | 3373 |
| Harry Bridges Plaza (Ferry Building) | 3163 |
| Market at 4th | 1869 |
| Powell Street BART | 1672 |
| San Francisco Caltrain (Townsend at 4th) | 1661 |
| 2nd at Townsend | 1580 |
| Embarcadero at Bryant | 1335 |
| Steuart at Market | 1219 |
| Grant Avenue at Columbus Avenue | 1098 |
| Market at Sansome | 1078 |

### Overall & Average Bike Usage

The additional trips are roughly uniformly distributed across the months as can be seen in how while the average individual bike usage is elevated, closer to 2% than 1.2%, the overall shape of the distribution is the same as Figure 25.



Figure 32. Average bike usage per month for individual bikes

The top 20 bikes used during the month vary from the restricted case. This may indicate bikes that may be in greater need of repair given the longer durations of trips they have been used in.

Table 11: Top 20 bikes (IDs listed) used per month

| Rank | Jan. | Feb. | Mar. | Apr. | May | Jun. | Jul. | Aug. | Sept. | Oct. | Nov. | Dec. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 661 | 285 | 4 | 342 | 638 | 601 | 228 | 80 | 653 | 660 | 582 | 355 |
| 2 | 509 | 294 | 206 | 602 | 118 | 164 | 5 | 438 | 604 | 388 | 410 | 541 |
| 3 | 314 | 630 | 415 | 542 | 377 | 134 | 454 | 579 | 175 | 455 | 325 | 452 |
| 4 | 61 | 76 | 662 | 513 | 482 | 382 | 113 | 676 | 422 | 228 | 424 | 75 |
| 5 | 55 | 419 | 514 | 239 | 370 | 85 | 251 | 309 | 670 | 474 | 563 | 250 |
| 6 | 553 | 543 | 233 | 531 | 495 | 124 | 598 | 319 | 15 | 531 | 653 | 344 |
| 7 | 657 | 243 | 357 | 495 | 345 | 394 | 395 | 343 | 174 | 265 | 416 | 571 |
| 8 | 534 | 52 | 566 | 256 | 421 | 654 | 297 | 333 | 188 | 323 | 586 | 550 |
| 9 | 363 | 386 | 33 | 318 | 643 | 90 | 524 | 243 | 21 | 644 | 488 | 379 |
| 10 | 561 | 226 | 442 | 208 | 351 | 33 | 572 | 172 | 393 | 316 | 539 | 414 |
| 11 | 517 | 350 | 332 | 450 | 398 | 289 | 346 | 598 | 304 | 372 | 487 | 570 |
| 12 | 449 | 452 | 219 | 357 | 286 | 348 | 548 | 538 | 489 | 436 | 451 | 467 |
| 13 | 345 | 584 | 450 | 582 | 30 | 352 | 467 | 308 | 76 | 67 | 343 | 476 |
| 14 | 596 | 591 | 354 | 126 | 598 | 531 | 398 | 8 | 518 | 72 | 420 | 390 |
| 15 | 246 | 552 | 552 | 337 | 593 | 281 | 252 | 586 | 368 | 521 | 246 | 330 |
| 16 | 533 | 307 | 600 | 500 | 335 | 431 | 416 | 354 | 370 | 347 | 379 | 448 |
| 17 | 343 | 418 | 525 | 474 | 575 | 373 | 600 | 327 | 459 | 600 | 523 | 435 |
| 18 | 291 | 425 | 476 | 293 | 521 | 380 | 497 | 582 | 583 | 627 | 427 | 558 |
| 19 | 301 | 518 | 437 | 397 | 310 | 539 | 594 | 541 | 496 | 76 | 99 | 508 |
| 20 | 324 | 476 | 560 | 549 | 379 | 598 | 511 | 384 | 392 | 567 | 304 | 363 |

### Trip & Weather Correlation

Similar correlation results were obtained in the restricted case as were obtained in the expanded case.



Figure 33. Correlation coefficients for weather and trip variables for all users



Figure 34. Correlation coefficients for weather and trip variables for all customers



Figure 35. Correlation coefficients for weather and trip variables for subscribers



Figure 36. Correlation coefficients for weather and city variables for all users

# Next Steps

Some of the next steps in the analysis could be to examine the weather imputation. The stations are localized into 3 clusters (Figure 39), and so it may be possible to impute weather for nearby stations depending upon their proximity.



Figure 37. Latitude and longitude coordinates for individual stations

It will also be important to further explore the differences between subscribers and customers. There is already some evidence that subscribers and customers are separated by behaviour, as is demonstrated in Table 6. While the correlations between the groups vary, the significance of those variations will require significance testing.

Furthermore, the actual modelling will require linear regression models and sensitivity analyses. It may also behoove the analysis to try and gain additional behavioural data, which are hidden confounding factors in how customers and subscribers are influenced to decide to bike, or for how long to bike. A worker who subscribes for their bike may ignore the weather due to the necessity to reach their workplace, while a tourist customer may try to use the bike regardless of the weather due to the limited time they have.

Ultimately, an understanding of who is using the bikes, where they are going, when they are going, and how often they are going there will aid in developing repair schedules for the bikes and stations.

# References

1. Climate of San Francisco narrative. (2024, April 14). https://ggweather.com/sf/daily.html
2. The Weather Channel. (2018, January 11). UK fog risk: What is the difference mist, smog, haze and fog? https://weather.com/en-GB/unitedkingdom/weather/news/2018-01-09-uk-weather-forecast-what-is-fog-mist-cloud-moisture-humidity
3. How we measure visibility. Met Office. (n.d.). https://www.metoffice.gov.uk/weather/guides/observations/how-we-measure-visibility#:~:text=Visibility%20sensors%20measure%20the%20meteorological,5%25%20of%20its%20original%20value.
4. US Department of Commerce, N. (2017, June 20). NWS Lix - Hurricane Wind Threats and impacts. National Weather Service. https://www.weather.gov/lix/htiwind#:~:text=Hurricane%20winds%2074%20to%2090,of%20mobile%20homes%20is%20likely.
5. 511. Metropolitan Transportation Commission. (2021, July 7). https://mtc.ca.gov/operations/traveler-services/511
6. Edward Epstein, C. S. W. (2000, March 2). New way to the wharf / merchants hope F-market line will draw locals to tourist attractions. San Francisco Chronicle. https://www.sfchronicle.com/news/article/New-Way-to-the-Wharf-Merchants-hope-F-Market-3304845.php
7. Harry Bridges plaza - plaza - phone number - hours - photos - embarcadero. SF Station. (n.d.). https://www.sfstation.com/harry-bridges-plaza-b7616