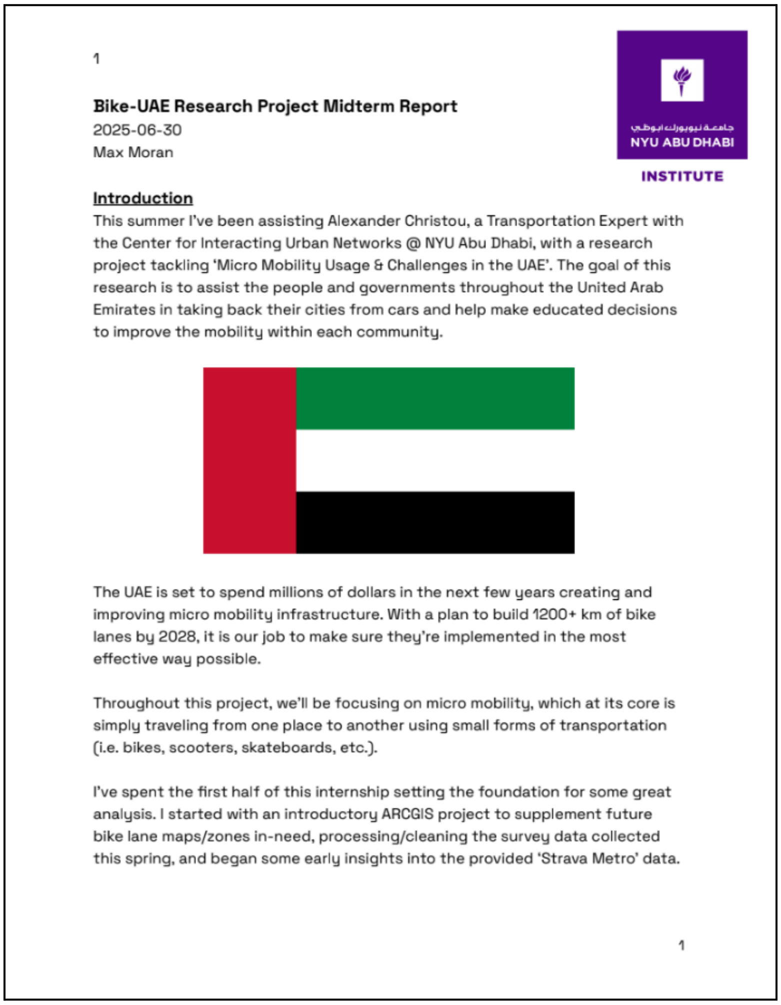
Bike-UAE Research Project Final Report

2025-07-31
Max Moran

Introduction

This summer I worked with Alexander Cristou to research and understand what encourages cycling in the United Arab Emirates; with a goal of making biking, and other forms of micromobility, not only safe but also enjoyable.

The first part of this project focused around understanding and cleaning the complex survey dataset provided to me. With 100+ variables in different languages, formats, and contexts it was important to create a usable base for analysis. You can review that documentation below as needed:



1

Bike-UAE Research Project Midterm Report
2025-06-30
Max Moran

جامعة نيويورك ابوظبي
NYU ABU DHABI
INSTITUTE

Introduction
This summer I've been assisting Alexander Christou, a Transportation Expert with the Center for Interacting Urban Networks @ NYU Abu Dhabi, with a research project tackling 'Micro Mobility Usage & Challenges in the UAE'. The goal of this research is to assist the people and governments throughout the United Arab Emirates in taking back their cities from cars and help make educated decisions to improve the mobility within each community.

The UAE is set to spend millions of dollars in the next few years creating and improving micro mobility infrastructure. With a plan to build 1200+ km of bike lanes by 2028, it is our job to make sure they're implemented in the most effective way possible.

Throughout this project, we'll be focusing on micro mobility, which at its core is simply traveling from one place to another using small forms of transportation (i.e. bikes, scooters, skateboards, etc.).

I've spent the first half of this internship setting the foundation for some great analysis. I started with an introductory ARCGIS project to supplement future bike lane maps/zones in-need, processing/cleaning the survey data collected this spring, and began some early insights into the provided 'Strava Metro' data.

1

Bike-UAE Midterm Report

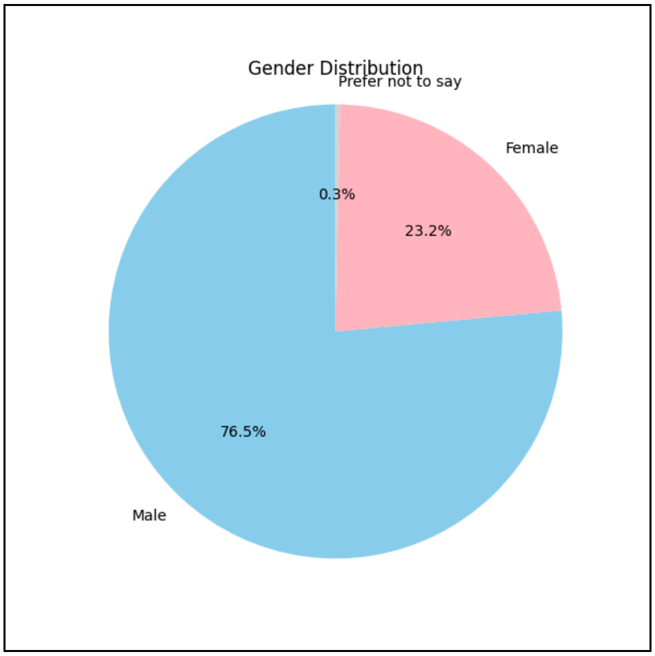
The Goal

These past few weeks I've been focused on understanding the data and finding solutions. The first step in this process was to establish what exactly we are trying to test/solve. The overall idea is to find what gets people out riding, regardless of the method of micromobility.

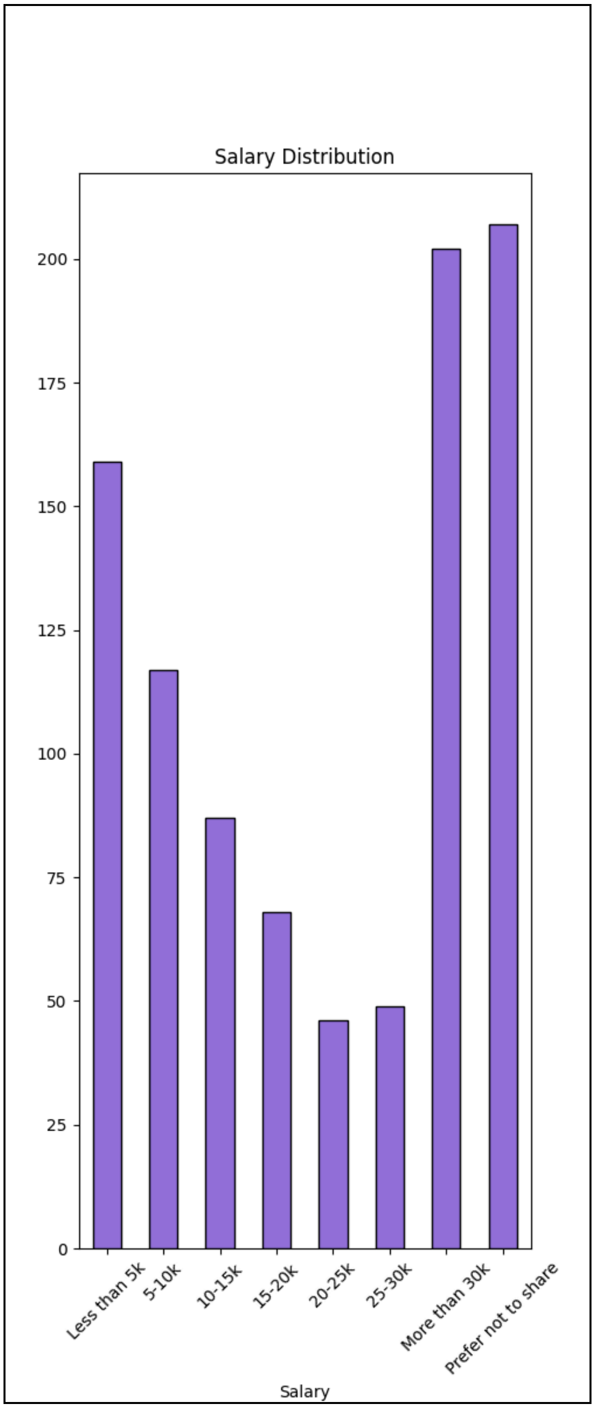
With a good set of questions asked in the survey, we have a pretty solid idea of what factors should lead to increased ridership. Using Python, I attempted to prove what gets people on the bike. You can find my code for my project along with other helpful files in the public GitHub here.

The People

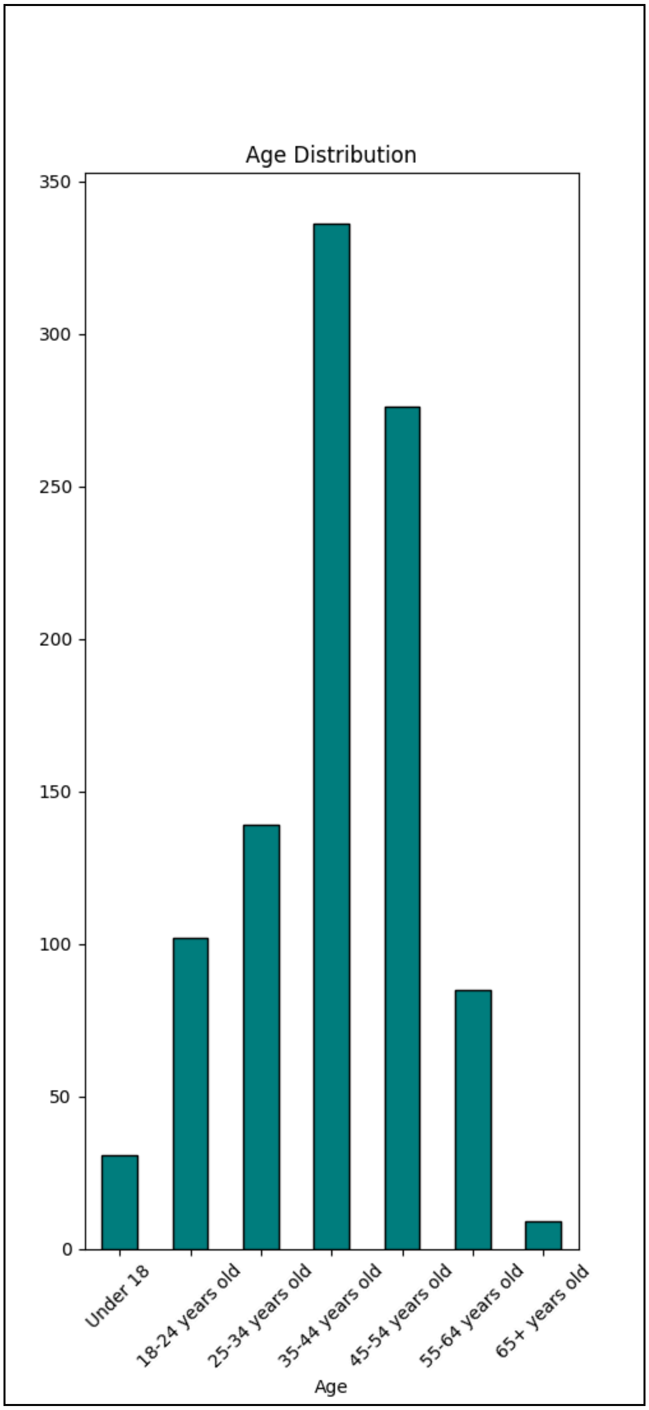
Before trying to prove why these people do/don't ride, let's establish who they are. Below you'll find some visualizations of what kind of people filled out this survey:



Gender Distribution

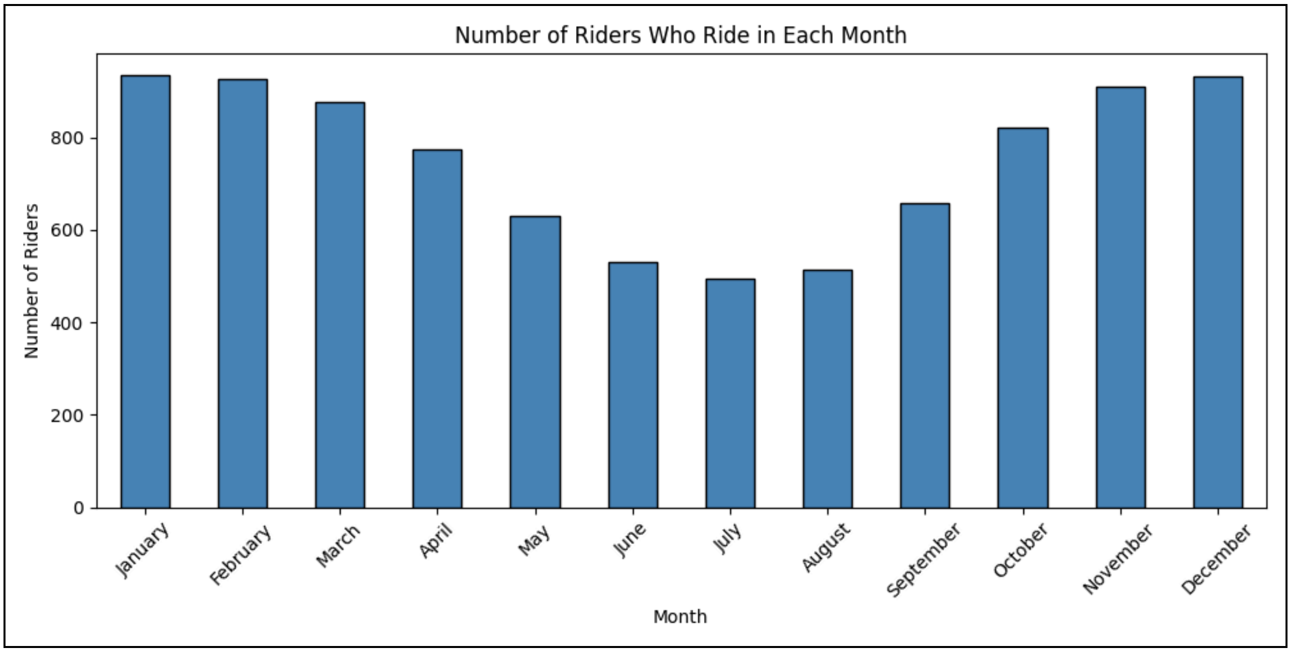


Salary Distribution

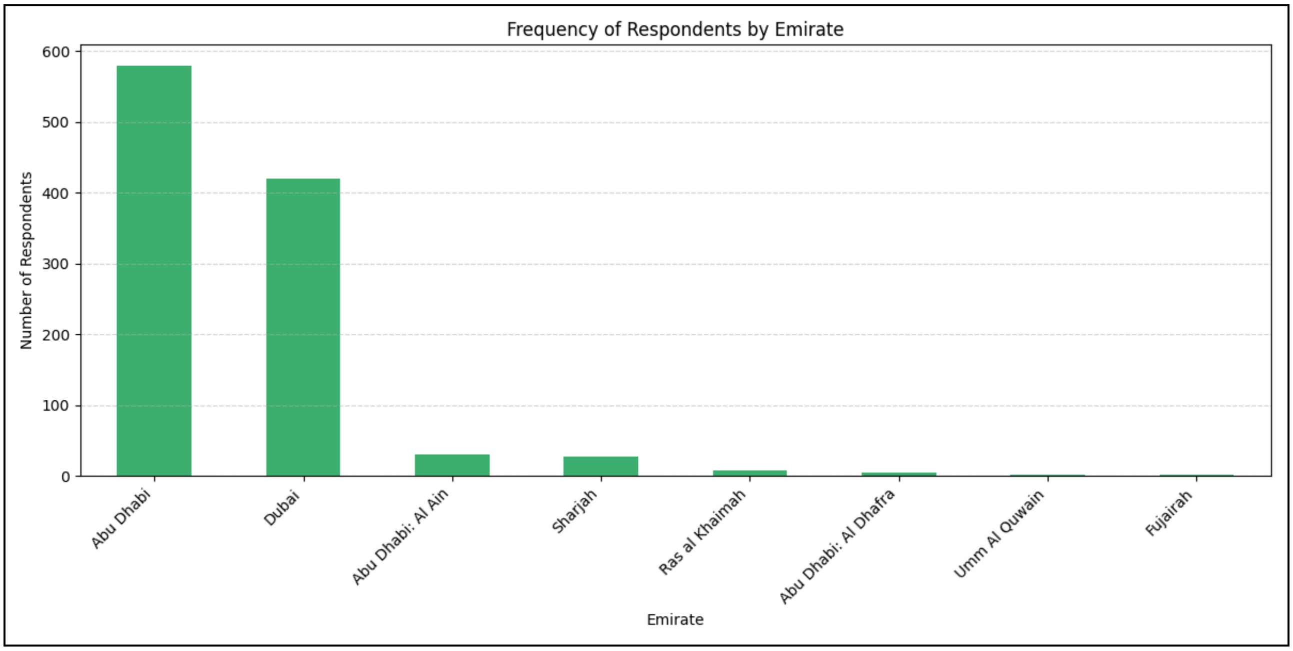


Age Distribution

We can see from the charts above that we're dealing with mostly middle-aged men who are either doing very well financially or not so much. This tracks a lot with the cycling population you see in the United States, with the sport attracting both the rich and the poor.



Number of Riders in Each Month



Number of Riders in Each Emirate

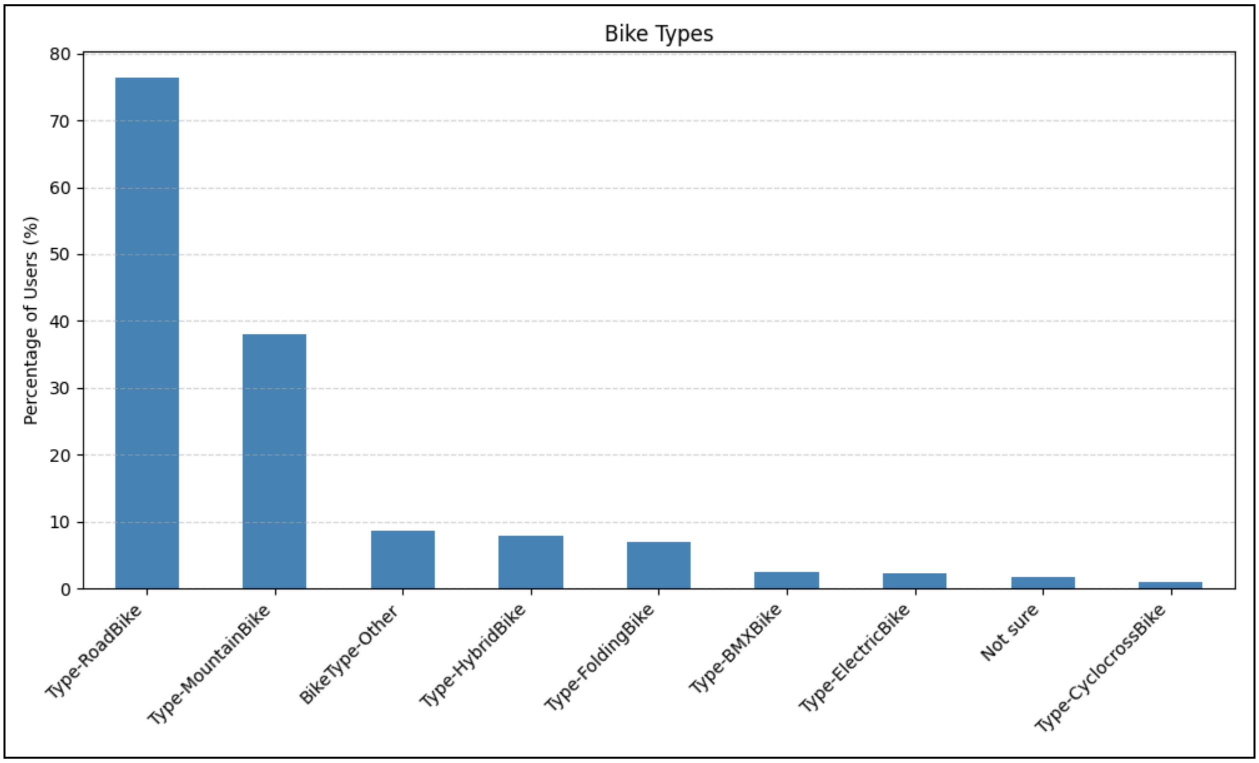
Neither of these charts are surprising at all. The majority of the population lives in Abu Dhabi or Dubai due to a number of factors, so it only makes sense we'd see that distribution here. The UAE also has some of the hottest summers in the world so a dip in cycling is also expected during the summer months.

This should be sufficient in giving us a decent picture of who we are dealing with. We can also see in the data a large proportion of individuals hailing from the Philippines and India, both countries with different expectations for transport compared to the West.

We're almost done laying the foundation for some great analysis. The last things to clear up are text-entry responses from the user and translating Arabic entries.

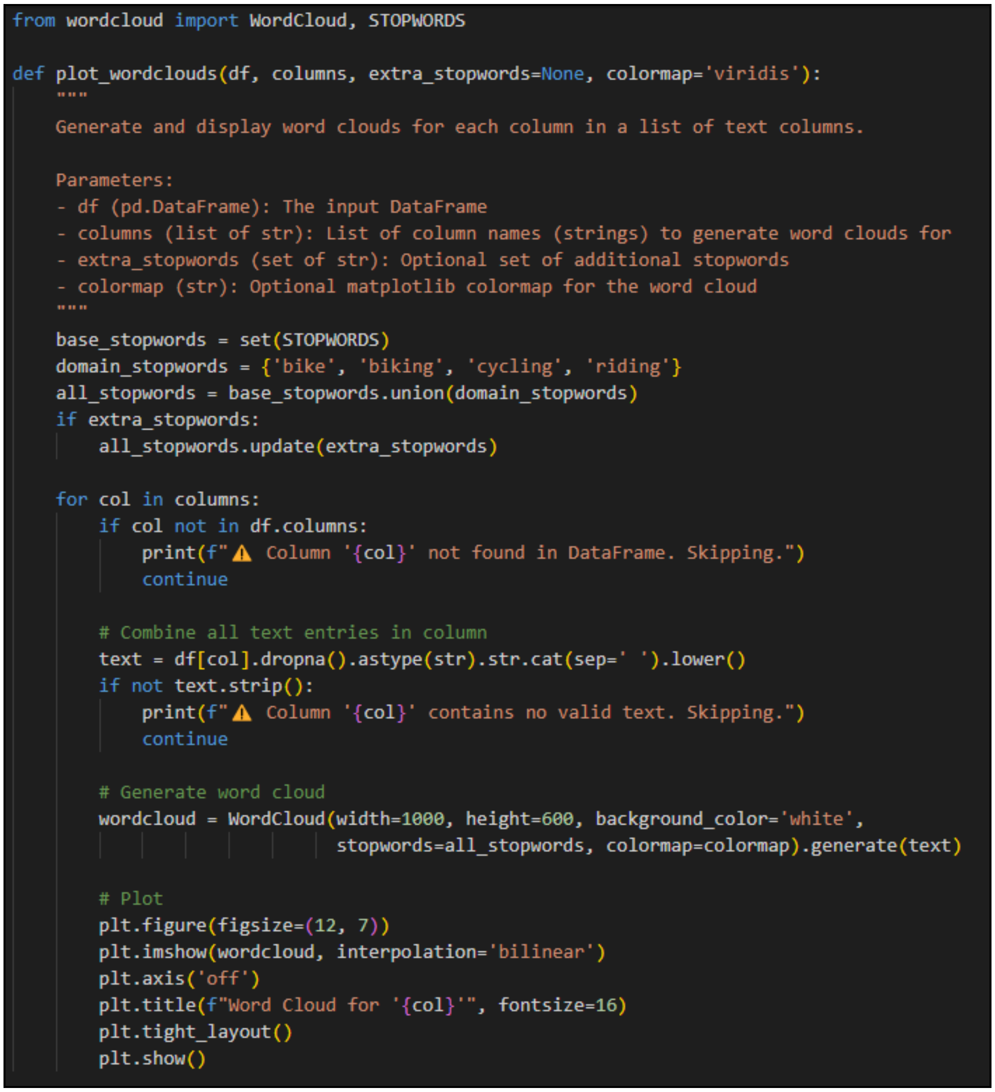
Wordclouds

Throughout the survey there were plenty of options for user text entry. In many cases this was to fill in the 'Other' option which sometimes was a large proportion of answers.



Bike Types

To help sort through these and get a grasp of what people care about I created word clouds of each of these sections to see what was commonly repeated for these questions.

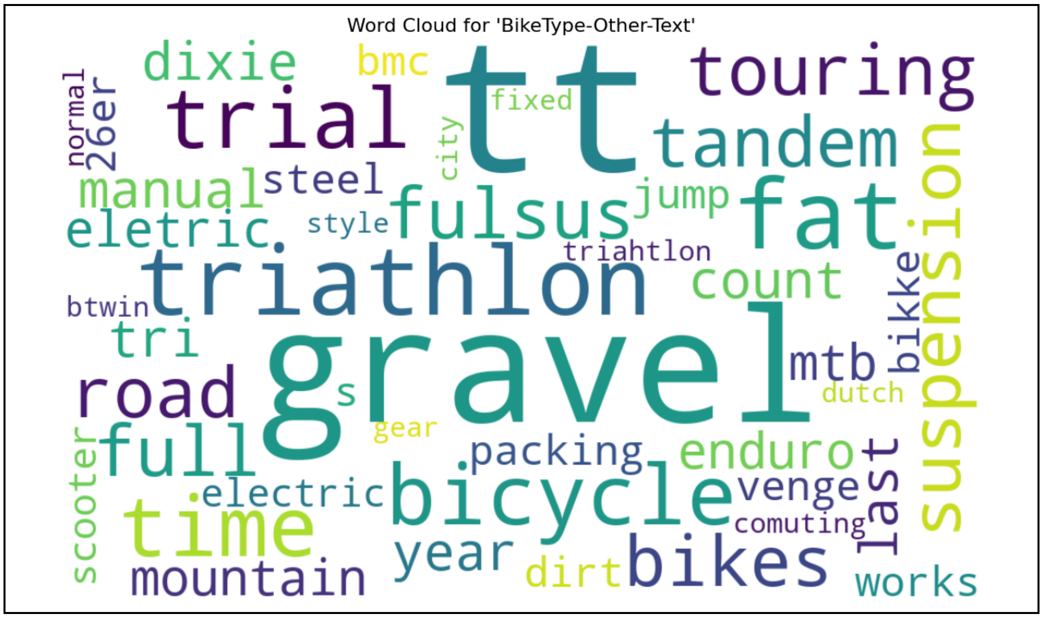


```python
from wordcloud import WordCloud, STOPWORDS

def plot_wordclouds(df, columns, extra_stopwords=None, colormap='viridis'):
    """
    Generate and display word clouds for each column in a list of text columns.

    Parameters:
    - df (pd.DataFrame): The input DataFrame
    - columns (list of str): List of column names (strings) to generate word clouds for
    - extra_stopwords (set of str): Optional set of additional stopwords
    - colormap (str): Optional matplotlib colormap for the word cloud
    """
    base_stopwords = set(STOPWORDS)
    domain_stopwords = {'bike', 'biking', 'cycling', 'riding'}
    all_stopwords = base_stopwords.union(domain_stopwords)
    if extra_stopwords:
        all_stopwords.update(extra_stopwords)

    for col in columns:
        if col not in df.columns:
            print(f"⚠ Column '{col}' not found in DataFrame. Skipping.")
            continue

        # Combine all text entries in column
        text = df[col].dropna().astype(str).str.cat(sep=' ').lower()
        if not text.strip():
            print(f"⚠ Column '{col}' contains no valid text. Skipping.")
            continue

        # Generate word cloud
        wordcloud = WordCloud(width=1000, height=600, background_color='white',
                              stopwords=all_stopwords, colormap=colormap).generate(text)

        # Plot
        plt.figure(figsize=(12, 7))
        plt.imshow(wordcloud, interpolation='bilinear')
        plt.axis('off')
        plt.title(f"Word Cloud for '{col}'", fontsize=16)
        plt.tight_layout()
        plt.show()
```

Wordcloud Code

This was a really cool way to see the data. One of my favorite aspects of good analysis is enjoyable data visualization. I could, of course, just show common words by frequency, but where's the fun in that?

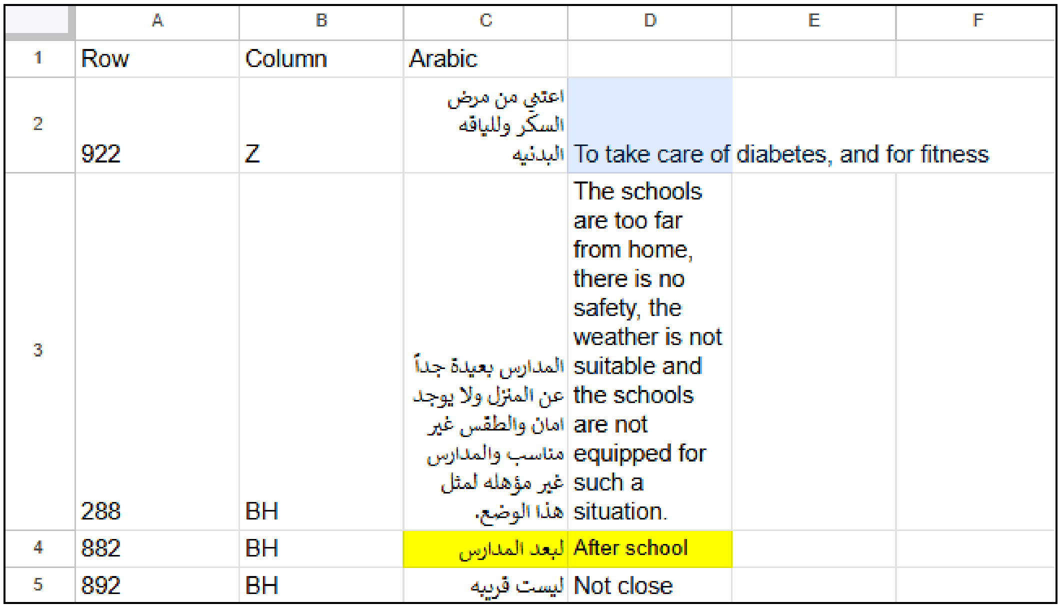
I haven't included all of the word clouds to avoid being excessive in this report, but if you'd like to review them you can find the results here. I've included, however, a couple along with the relevant result of different styles of bikes owned, which you can find on the next page.



Other Bike Types Owned

Translations

A unique aspect of this project was working on a ton of non-English responses, specifically Arabic. With unique characters, sentence structure, and a right-to-left order this was a different sort of challenge I hadn't faced before. Thankfully Mr. Cristou is fluent and was able to help me translate the Arabic cells to the correct English conversion.

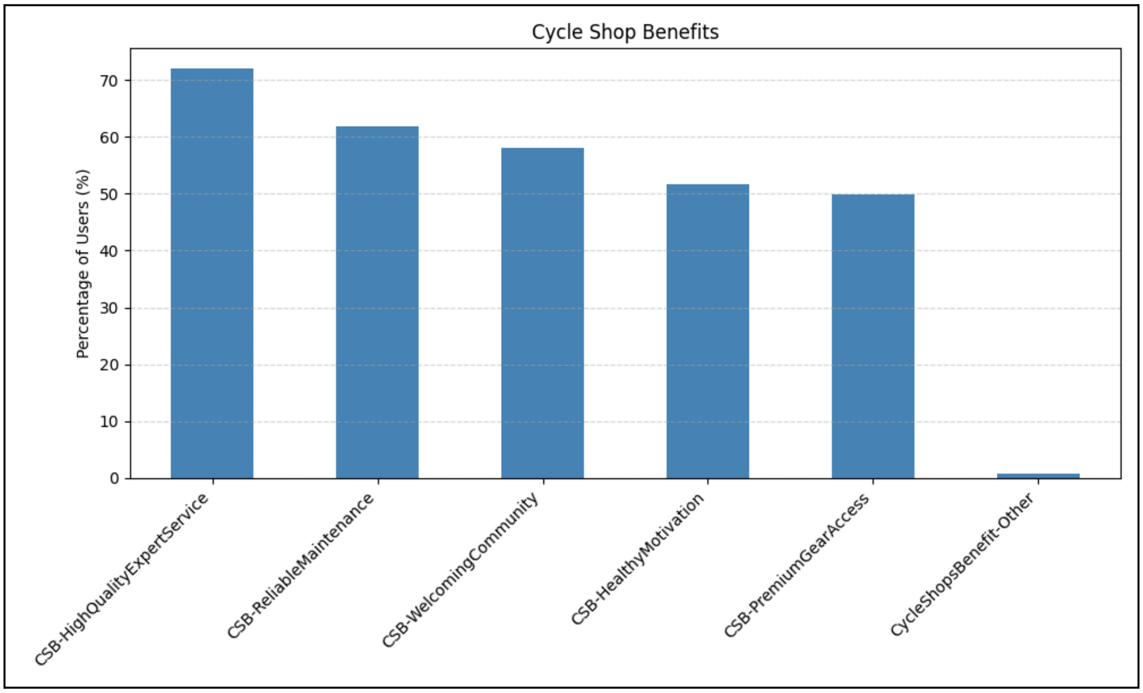


	A	B	C	D	E	F
1	Row	Column	Arabic			
2	922	Z	اعتني من مرض السكر وللياقه البدنيه	To take care of diabetes, and for fitness		
3	288	BH	المدارس بعيدة جداً عن المنزل ولا يوجد امان والطقس غير مناسب والمدارس غير مؤهله لمثل هذا الوضع.	The schools are too far from home, there is no safety, the weather is not suitable and the schools are not equipped for such a situation.		
4	882	BH	لبعد المدارس	After school		
5	892	BH	ليست قريبه	Not close		

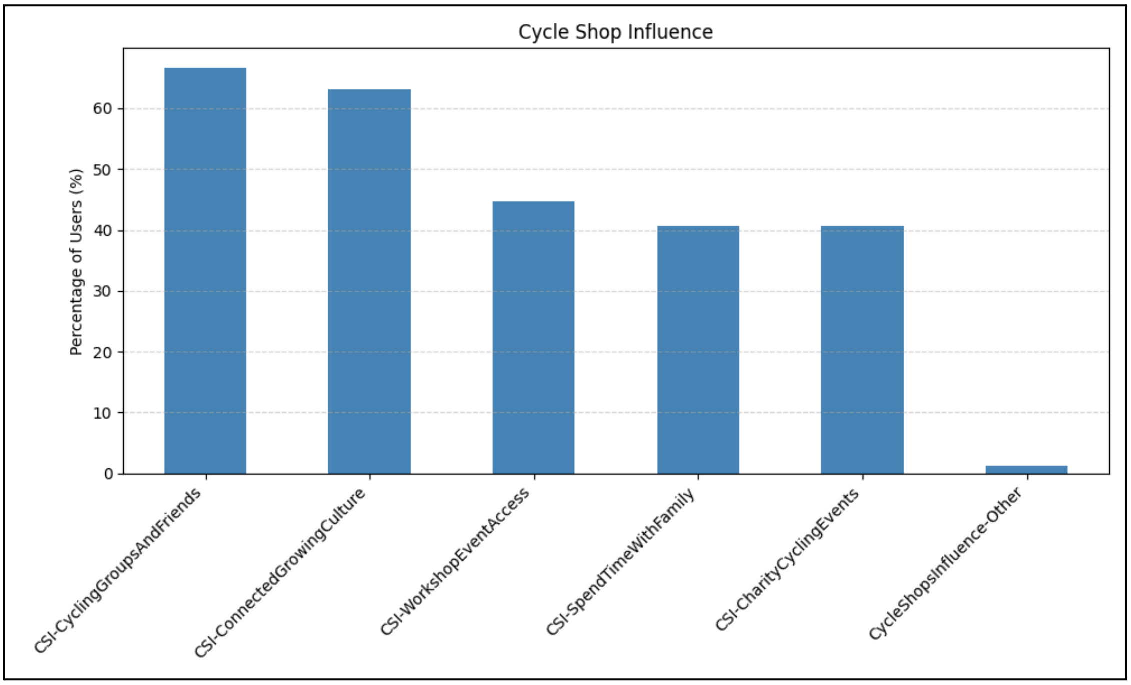
Sample of Arabic Translations

User Input

Let's take a look at what the riders care about, or at least how their responses reflect that. We'll take this section by section to make it more digestible. First let's cover how cycle shops affect riders in the UAE:



Cycle Shop Benefits



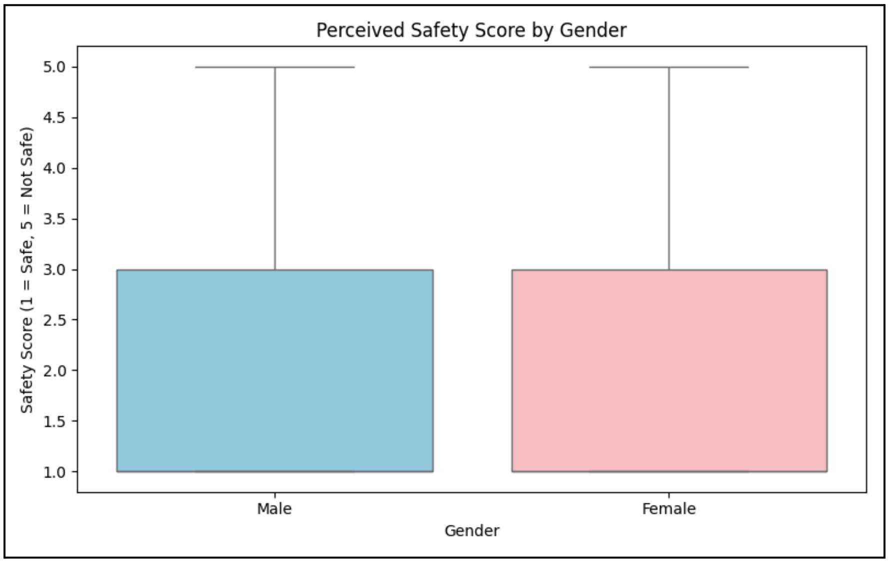
Cycle Shop Influences

From these charts you can see what people in the region find important about cycle shops in their Emirate. We can see that the skilled individuals that work there, as well as the community that naturally forms from establishing a biking hub, are what matter to these riders.

A theme we see throughout the data is the importance of humanity when it comes to riding. If you want to separate someone from society, you put them in a 3 ton steel box with tinted windows that moves 10x a walking pace. Bikes allow us to enjoy the environment around us, meeting people in the process.

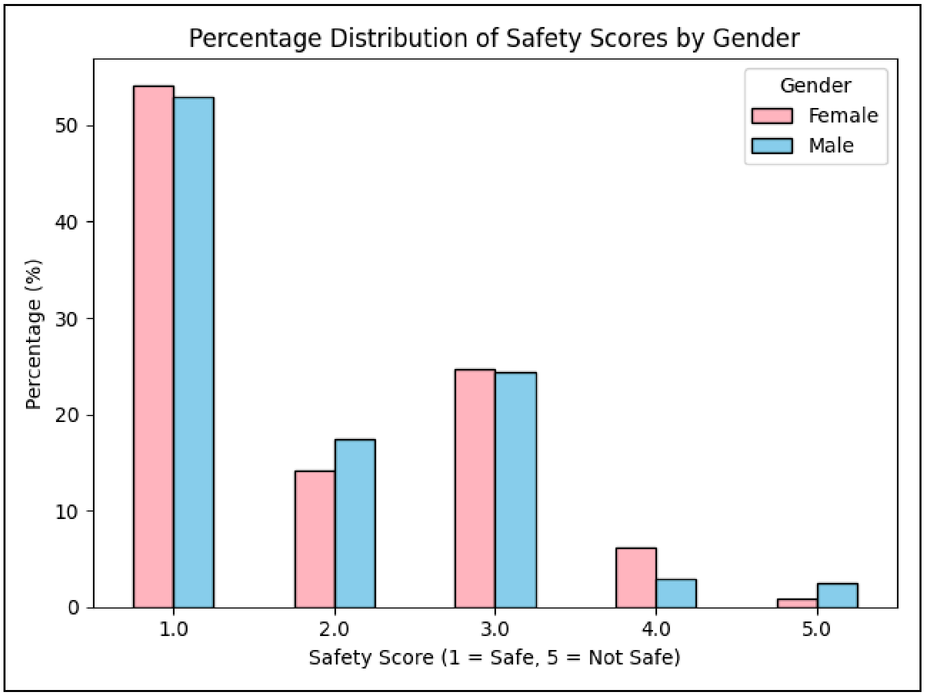
Gender

I have to admit I was guilty of some bias in this analysis. The UAE, and its surrounding countries, have had stories written about how women can be treated unfairly. When looking through the data, I fully expected to see stark differences in women and men with respect to biking in the UAE.



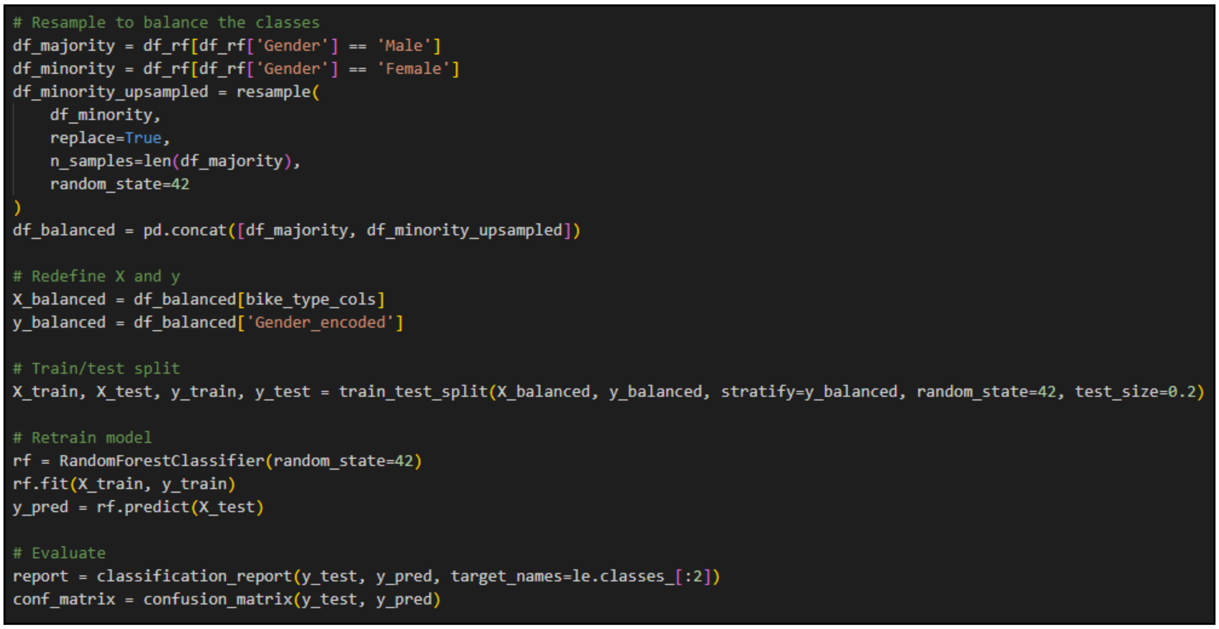
Safety Score by Gender

I tested the average safety score given by each gender to see if women felt more unsafe riding in their Emirates. But as you can see above there is virtually no difference. In the following plot I break it down by percentage within each score and the balance of men vs women is very similar.



Safety Score by Gender (%)

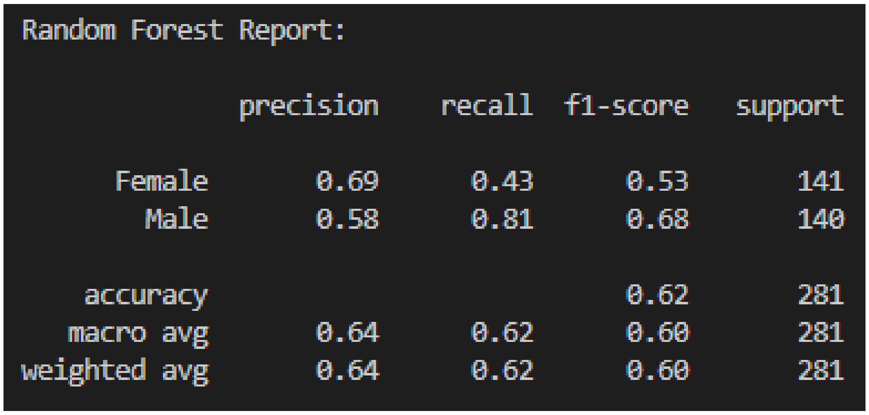
This shows why it is important to run these tests. I came in with the predisposed idea that women would feel more unsafe than men but I was totally wrong. I even created a random forest model, which splits data in an attempt to predict outcomes, to see if gender affects bike choice.



```
# Resample to balance the classes
df_majority = df_rf[df_rf['Gender'] == 'Male']
df_minority = df_rf[df_rf['Gender'] == 'Female']
df_minority_upsampled = resample(
    df_minority,
    replace=True,
    n_samples=len(df_majority),
    random_state=42
)
df_balanced = pd.concat([df_majority, df_minority_upsampled])

# Redefine X and y
X_balanced = df_balanced[bike_type_cols]
y_balanced = df_balanced['Gender_encoded']

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(X_balanced, y_balanced, stratify=y_balanced, random_state=42, test_size=0.2)

# Retrain model
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)

# Evaluate
report = classification_report(y_test, y_pred, target_names=le.classes_[:2])
conf_matrix = confusion_matrix(y_test, y_pred)
```

Random Forest Code

From the output report you can see how well the model could predict bike type by gender. This isn't necessarily a bad model but it isn't a great one.

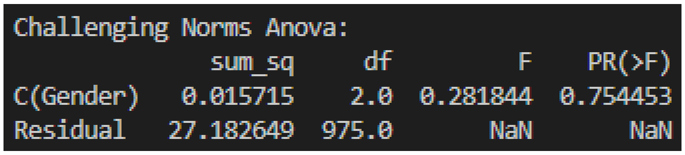


```
Random Forest Report:

              precision    recall  f1-score   support

      Female       0.69      0.43      0.53       141
        Male       0.58      0.81      0.68       140

    accuracy                           0.62       281
   macro avg       0.64      0.62      0.60       281
weighted avg       0.64      0.62      0.60       281
```

Random Forest Results

The f1-score shows basically how accurate your model is when it comes to False Negatives, False Positives, etc. From the results we can see it is only a little over half, meaning that maybe bike type and gender aren't correlated enough to warrant such a definitive yes or no.

I also looked to see if women were worried about challenging norms affecting their desire to go out and ride, and I was wrong again.



```
Challenging Norms Anova:
              sum_sq     df         F    PR(>F)
C(Gender)   0.015715    2.0  0.281844  0.754453
Residual   27.182649  975.0       NaN       NaN
```

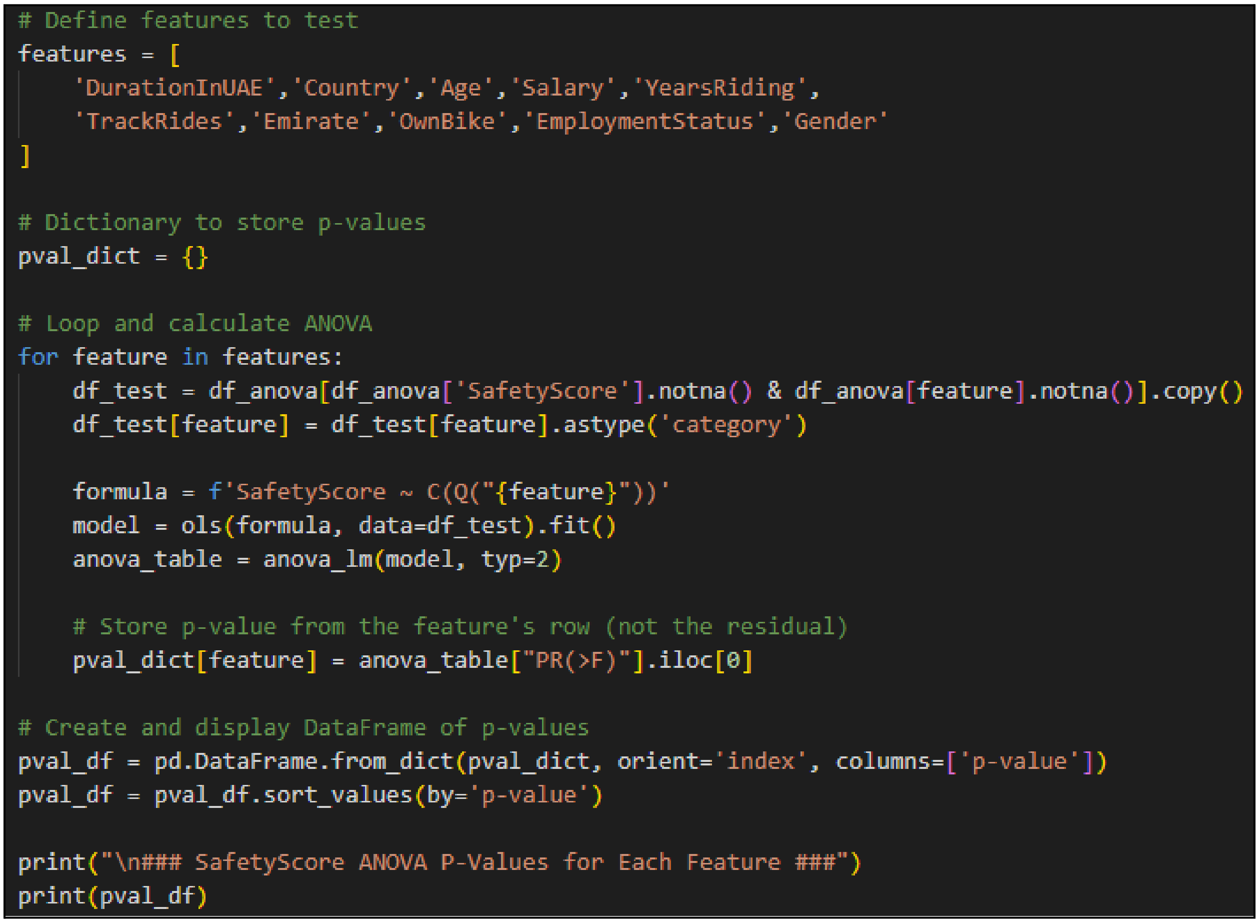
Challenging Norms ANOVA Results

With a p-value of 0.75, this suggests essentially no statistically significant evidence that gender and challenging norms being an issue are correlated.

It's important to acknowledge your biases before and during research. No one wants the answer that your subconscious manipulates into reality.

ANOVA

An important tool to see when features are statistically significant is the ANOVA table. This allows us to compare different features to see what is most important to a designated dependent variable.

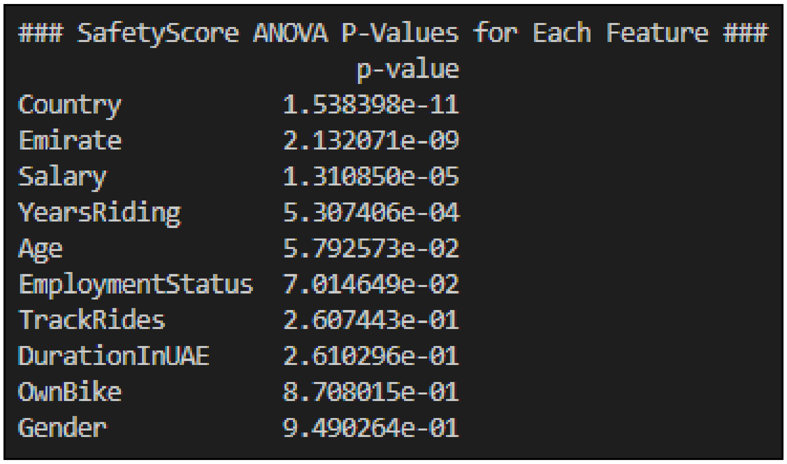


```
# Define features to test
features = [
    'DurationInUAE','Country','Age','Salary','YearsRiding',
    'TrackRides','Emirate','OwnBike','EmploymentStatus','Gender'
]

# Dictionary to store p-values
pval_dict = {}

# Loop and calculate ANOVA
for feature in features:
    df_test = df_anova[df_anova['SafetyScore'].notna() & df_anova[feature].notna()].copy()
    df_test[feature] = df_test[feature].astype('category')

    formula = f'SafetyScore ~ C(Q("{feature}"))'
    model = ols(formula, data=df_test).fit()
    anova_table = anova_lm(model, typ=2)

    # Store p-value from the feature's row (not the residual)
    pval_dict[feature] = anova_table["PR(>F)"].iloc[0]

# Create and display DataFrame of p-values
pval_df = pd.DataFrame.from_dict(pval_dict, orient='index', columns=['p-value'])
pval_df = pval_df.sort_values(by='p-value')

print("\n### SafetyScore ANOVA P-Values for Each Feature ###")
print(pval_df)
```

ANOVA Testing Code

With this python function, I was able to input different variables to see what most affected the output. For dependent variables I chose perceived safety and riding frequency. These felt like the best metrics for what we are trying to accomplish; get people out riding, safely.

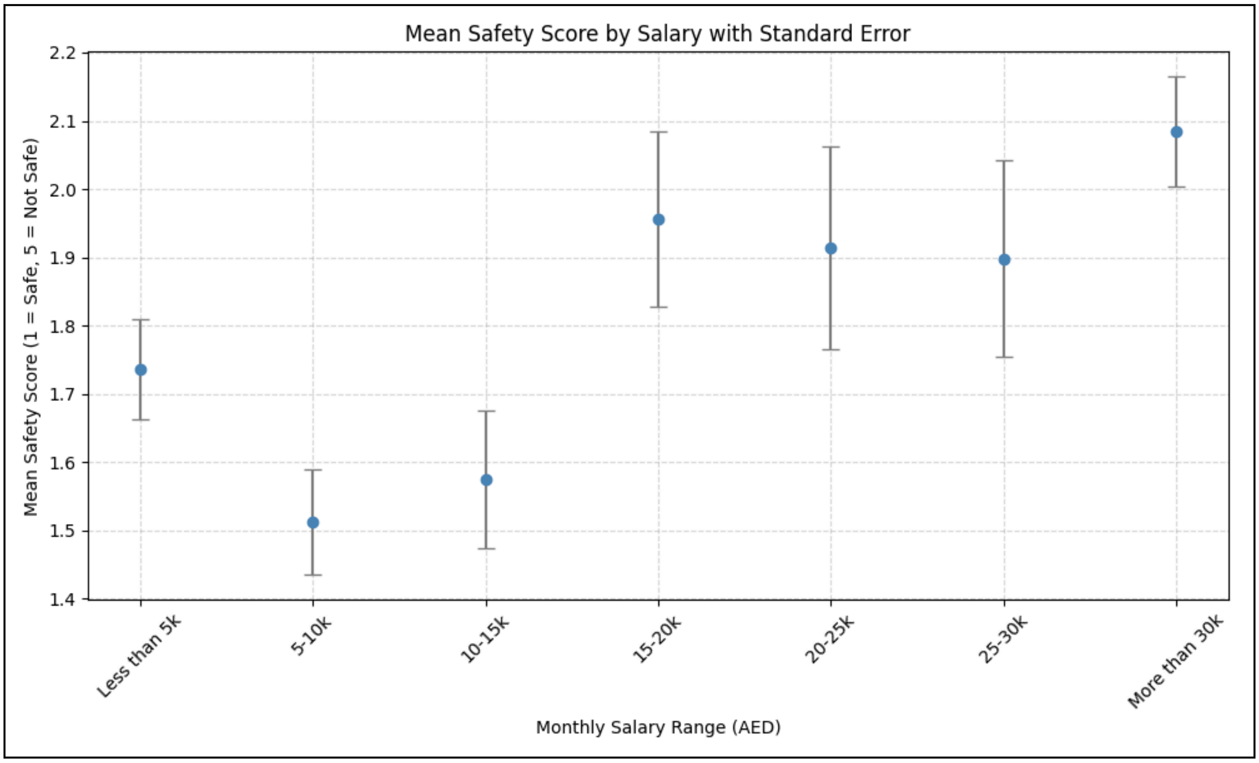
The output is below, with each feature ranked by importance essentially. The lower the p-value the more statistically significant it is.



```
### SafetyScore ANOVA P-Values for Each Feature ###
                       p-value
Country             1.538398e-11
Emirate             2.132071e-09
Salary              1.310850e-05
YearsRiding         5.307406e-04
Age                 5.792573e-02
EmploymentStatus    7.014649e-02
TrackRides          2.607443e-01
DurationInUAE       2.610296e-01
OwnBike             8.708015e-01
Gender              9.490264e-01
```

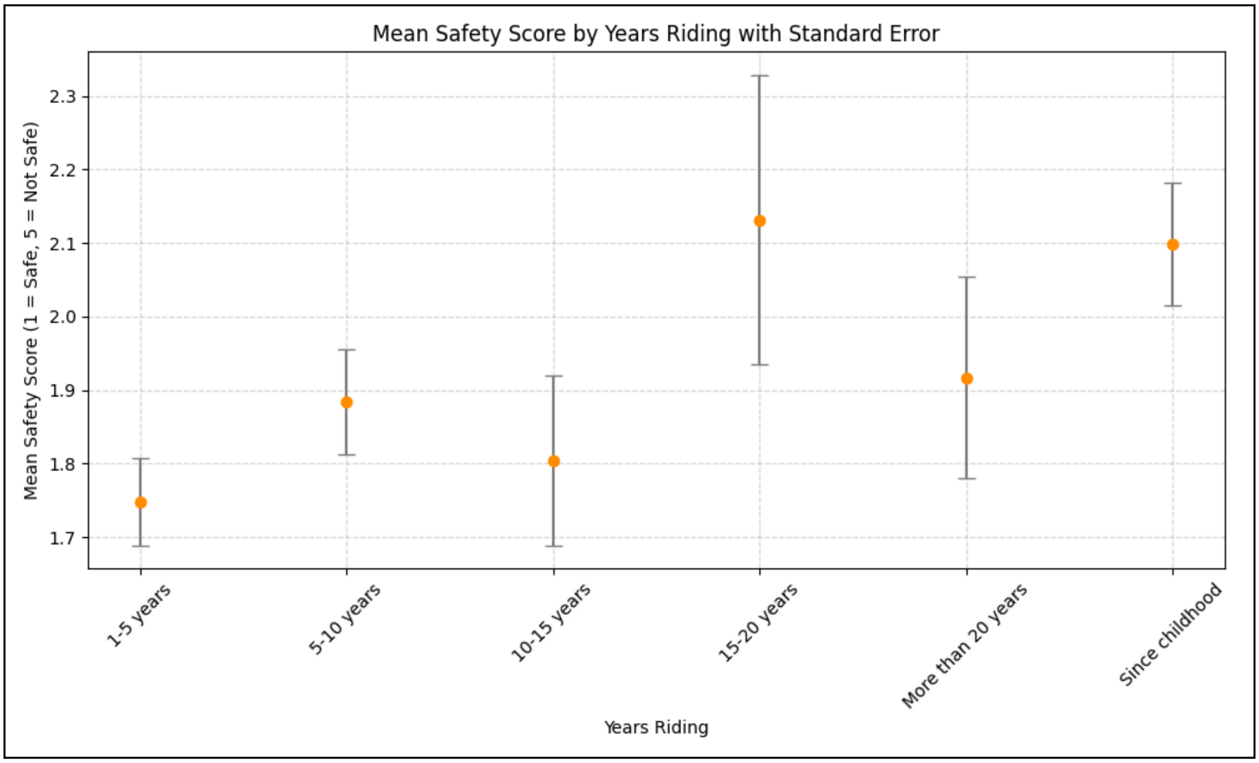
Safety Score ANOVA

For the perceived safety of the rider, you can see that some of the most important factors are country, emirate, salary, and years riding. Countries make a lot of sense, as different nationalities will have different expectations for what is acceptable as a micromobility path. Emirate also isn't a surprise due to the imbalance in biking infrastructure across the UAE. I plotted the mean safety score (with their SE) by salary and years riding to get a closer look.



Mean Safety Score by Salary

You can see in the chart above that as someone makes more money, they are more likely to feel unsafe on the roads. Perhaps this is due to those individuals having more to lose or having higher expectations of what the bare minimum safety standard should be.



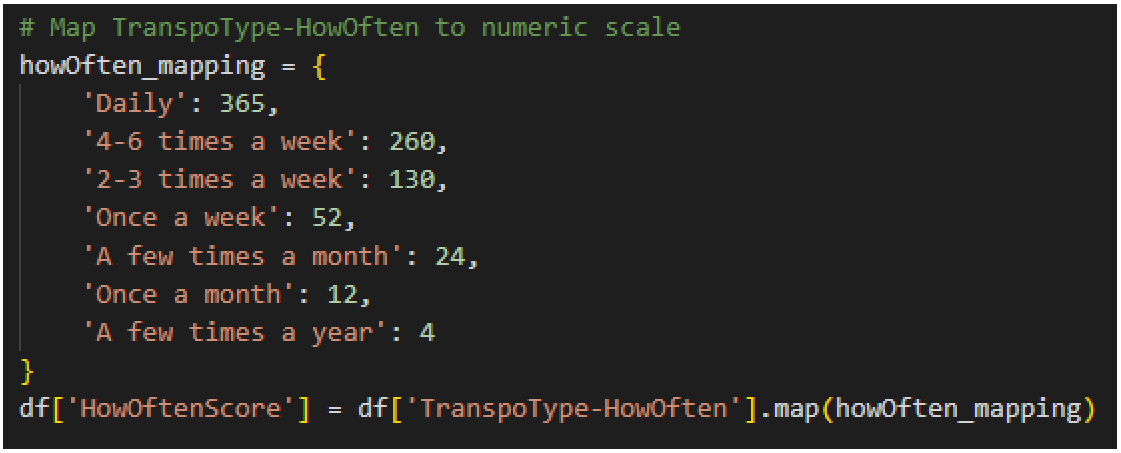
Mean Safety Score by Years Riding

Another interesting correlation is how the safety score trends upwards (less safe) as the rider has more experience. This seems counterintuitive, as obviously riders with more practice will be better in a wide range of scenarios.

I believe the reason for this correlation has a similar logic as the phrase, “Ignorance is bliss.” When you first start riding a bike, you stick to safe pathways with predictable traffic and infrastructure. You see the best of the best because you aren’t ready for anything else.

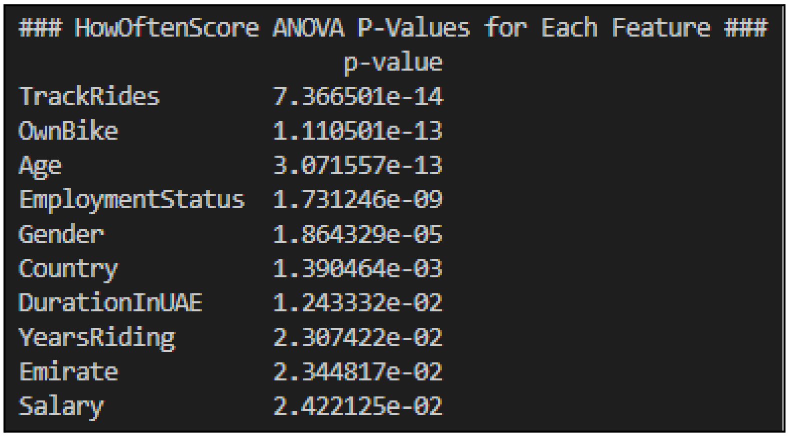
As you become a better rider, you want to experience what else your Emirate has to offer which quickly leads you into more hostile environments. It seems to me that the more you cycle, the more you see and realize how dangerous it can be.

Now for riding frequency, the riders were asked to select how often they rode. You can find the exact wording of the question/choices here along with the rest of the survey. Below you can see the ANOVA results for comparing leading features and the numerical conversion of how often people get out and ride.



```python
# Map TranspoType-HowOften to numeric scale
howOften_mapping = {
    'Daily': 365,
    '4-6 times a week': 260,
    '2-3 times a week': 130,
    'Once a week': 52,
    'A few times a month': 24,
    'Once a month': 12,
    'A few times a year': 4
}
df['HowOftenScore'] = df['TranspoType-HowOften'].map(howOften_mapping)
```

Categorial to Numeric Conversion for Riding Frequency



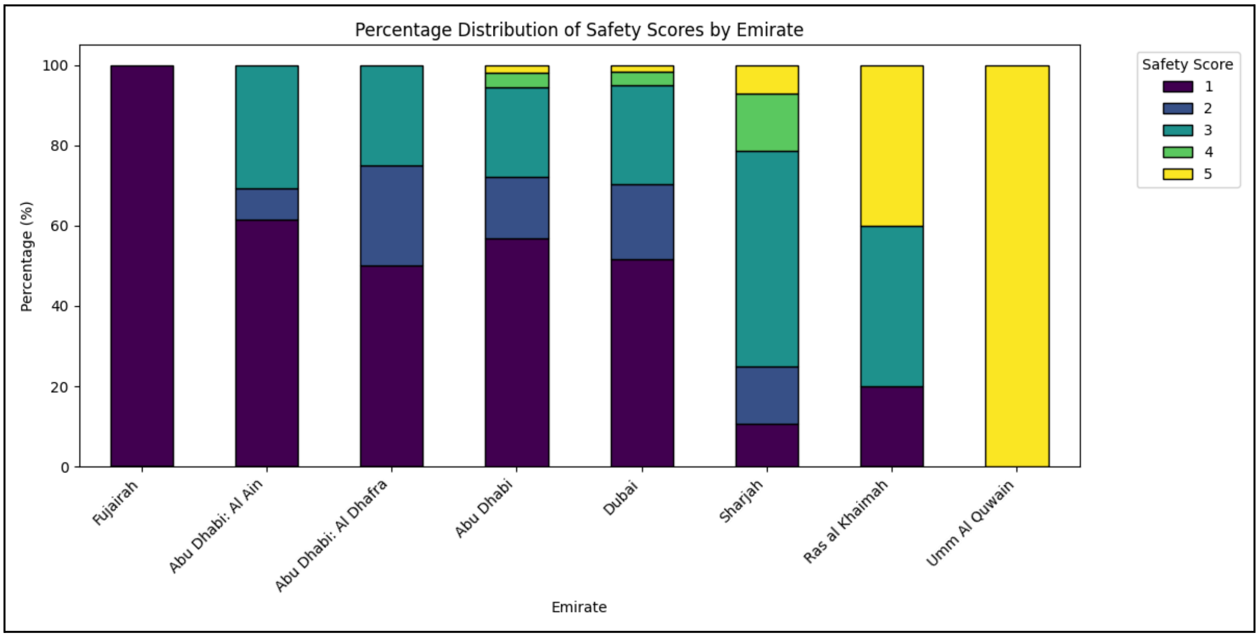
```
### HowOftenScore ANOVA P-Values for Each Feature ###
                        p-value
TrackRides         7.366501e-14
OwnBike            1.110501e-13
Age                3.071557e-13
EmploymentStatus   1.731246e-09
Gender             1.864329e-05
Country            1.390464e-03
DurationInUAE      1.243332e-02
YearsRiding        2.307422e-02
Emirate            2.344817e-02
Salary             2.422125e-02
```

How Often Score ANOVA

From the results the top features with the most importance were TrackRides, OwnBike, Age, and EmploymentStatus. TrackRides makes a lot of sense as services like Strava encourage the continual & repetitive use of their app for fitness, social interactions, and other reasons.

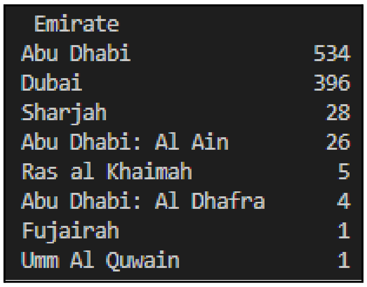
Owning a bike is also not a surprise feature, as having to rent a bike every time you want to go on a ride severely inhibits your ability to do so. Age makes sense as different ages affect your physical ability and stamina for repeat exercise.

I took the liberty of ranking the Emirates by safety as well to get an idea of how equal the UAE in this.



Emirate Ranked by Safety

I will say though that even though Fujairah and Umm Al Quwain are both the safest and unsafest, they both have just a single entry making them not the best source of evidence.

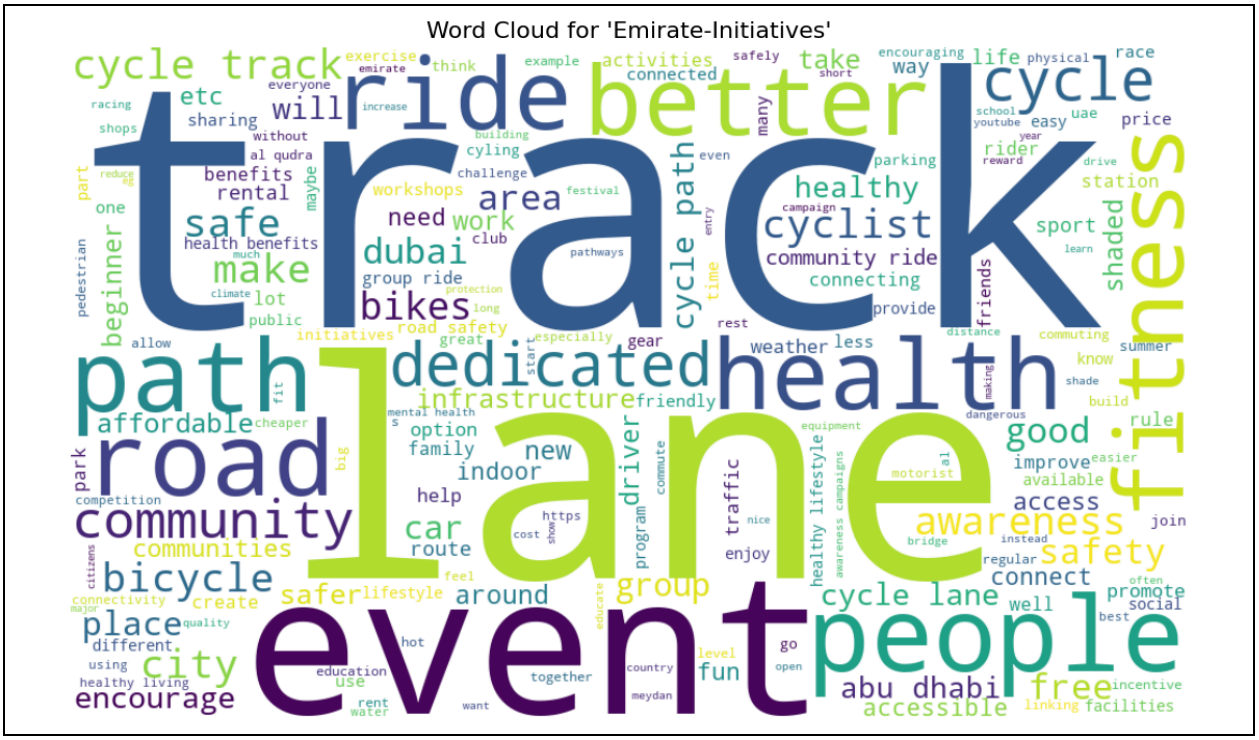


```
 Emirate
Abu Dhabi               534
Dubai                   396
Sharjah                  28
Abu Dhabi: Al Ain        26
Ras al Khaimah            5
Abu Dhabi: Al Dhafra      4
Fujairah                  1
Umm Al Quwain             1
```

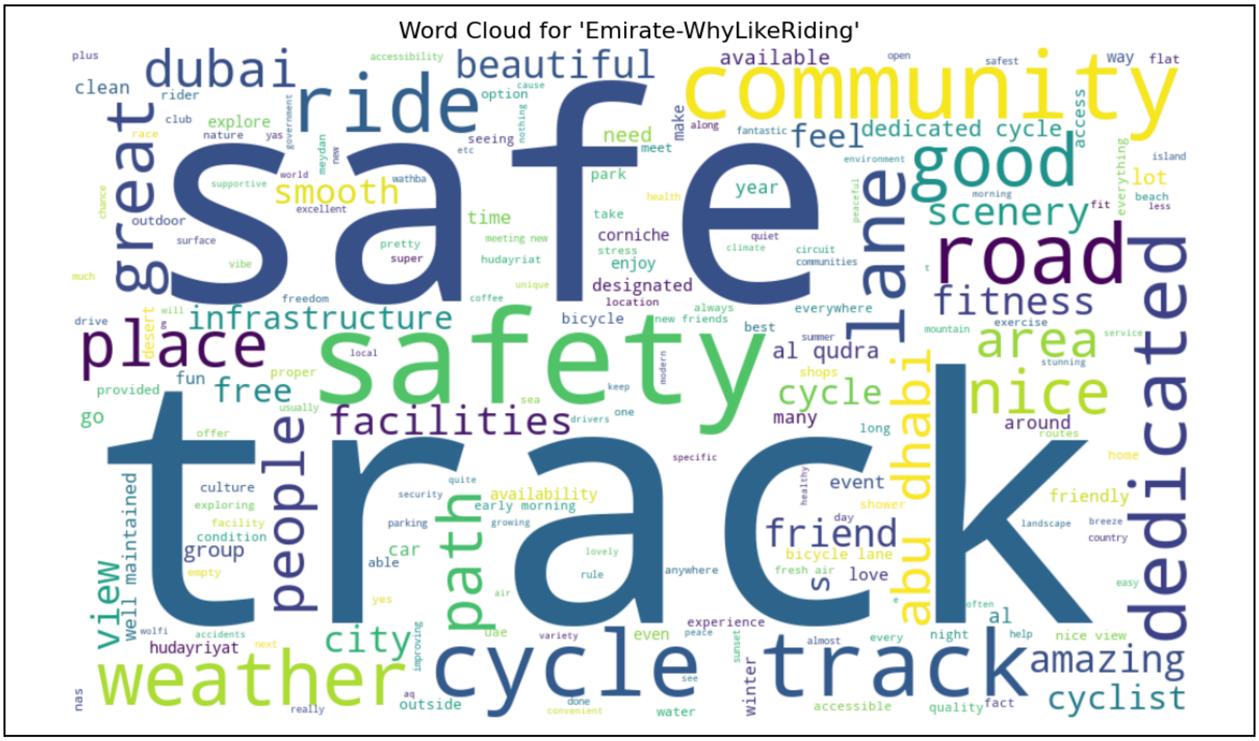
When considering the amount of entries per Emirate, the safest area within statistical significance is Abu Dhabi: Al Ain, with Sharjah having the lowest rated safety score on average.

Location

To emphasize the importance of bike paths, independent of cars, you only need to look at the word clouds associated with things the UAE does well or could do more of. We mostly see suggestions for safe tracks as well as recommendations for more community building events.

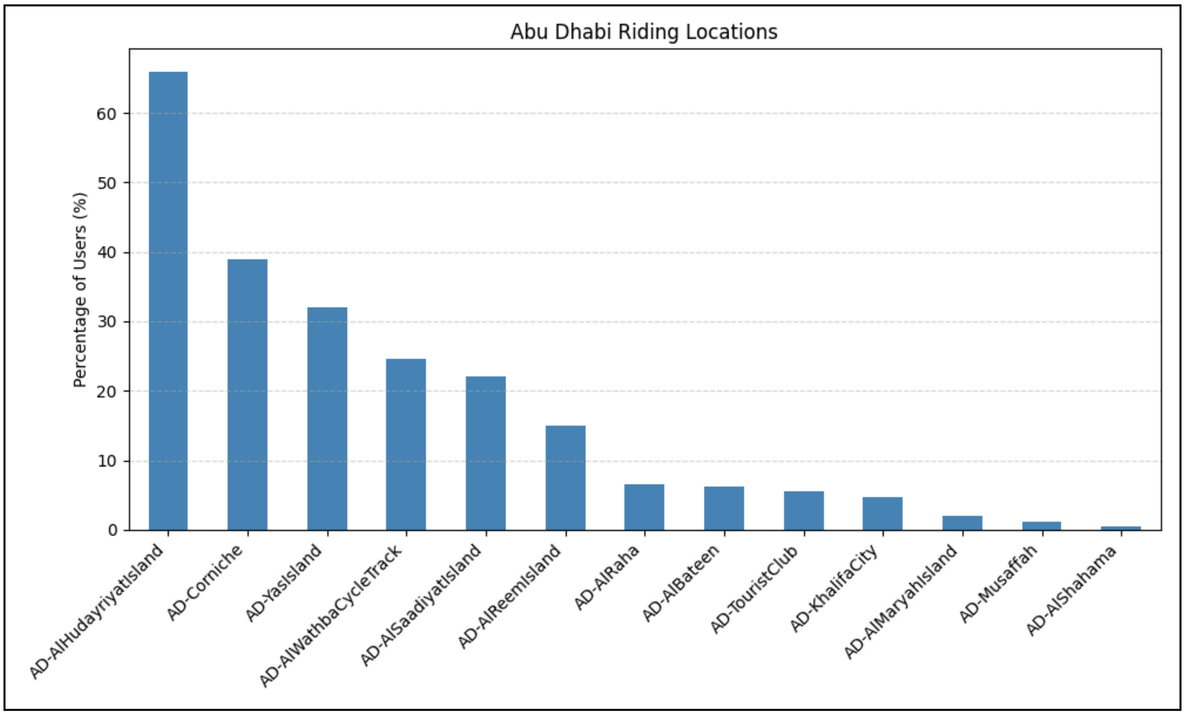


Initiatives encouraging more riding in their Emirate

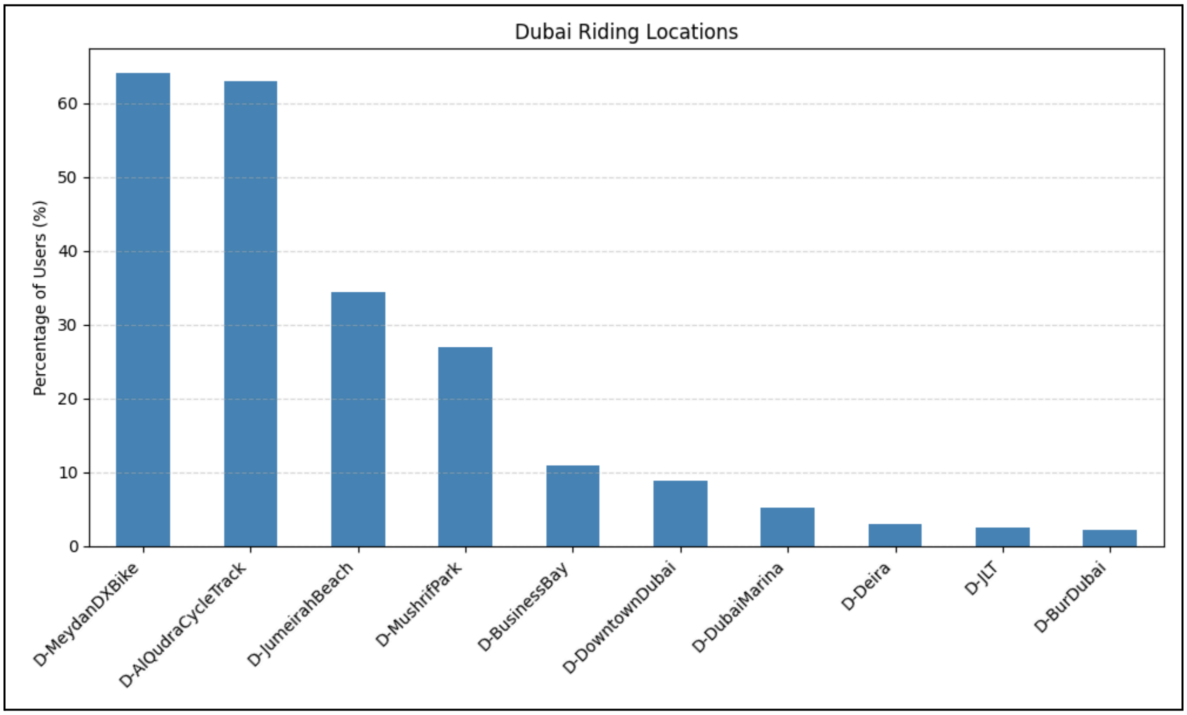


Why the user likes riding in their emirate

You can also see in the Emirate specific location responses that areas with protected bike paths usually get far more use than areas without.



Abu Dhabi Locations



Dubai Locations

The reason a lot of people don't ride bikes is due to a fear of getting hit by a car. The solution is in the data, showing that people ride where they feel safe. The goal of this project is to find what gets people to ride and we can clearly see with these plots that people want to and will ride in areas that are separated and protected from cars.

Results

After looking at all the evidence and comparing the different findings I have a good idea of what factors lead to getting people back on the road. You can see just from the survey demographics we should try to get more women involved in cycling across the UAE. A good solution to this would be creating 'Girls Night' style group rides to introduce new people to the sport.

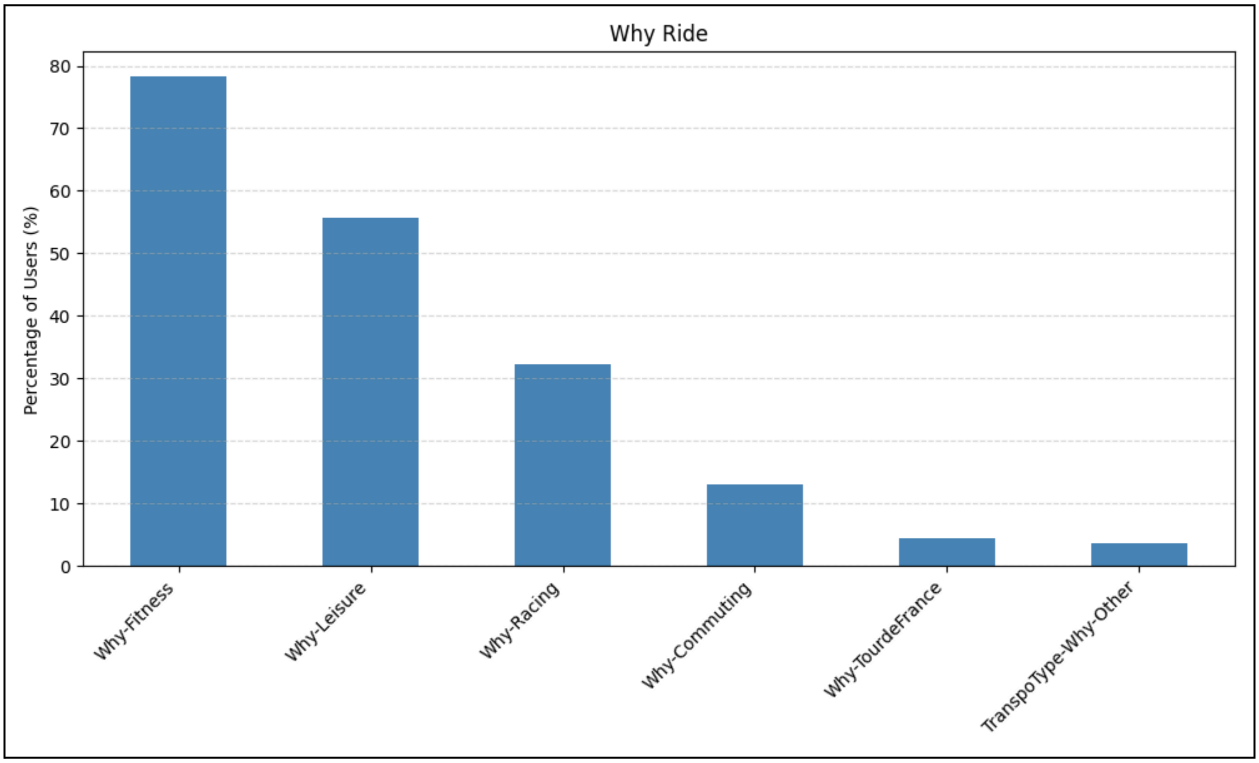
The extreme weather in the UAE was often cited as a reason for not riding, which isn't an easy problem to solve. A good starting solution would be to allow for shade installations along bike paths, in the form of trees or canopies. Keeping distances between destinations low would also help, lowering the time spent outside in the warmer months. This, like a lot of micromobility issues, come down to zoning and making areas favor density.

From my analysis, some of the most important factors to get people out riding are getting affordable and reliable bikes into the hands of those who need them. Owning a bike and tracking your rides statistically lead to a significant increase in ride frequency. On tracking rides, it can be hard to quantify community energy as a statistic, but the importance of simply sharing your rides with others can't be understated. Working with local governments or organizations like Strava to share community events and group rides really helps get people involved.

Lastly, it is clear from the analysis and word clouds that what people need the most is a safe place to bike. Overwhelmingly, bike tracks were favored over all other options. This shows how good infrastructure really is the key. I unfortunately ran out of time and was unable to get to any serious analysis of the Strava Metro data due to starting a new job and studying for my graduate degree comprehensive exams. However I still intend to help with that analysis as much as I can in August!

Love

My favorite result from this whole project comes from the question, “Why do you Ride?”

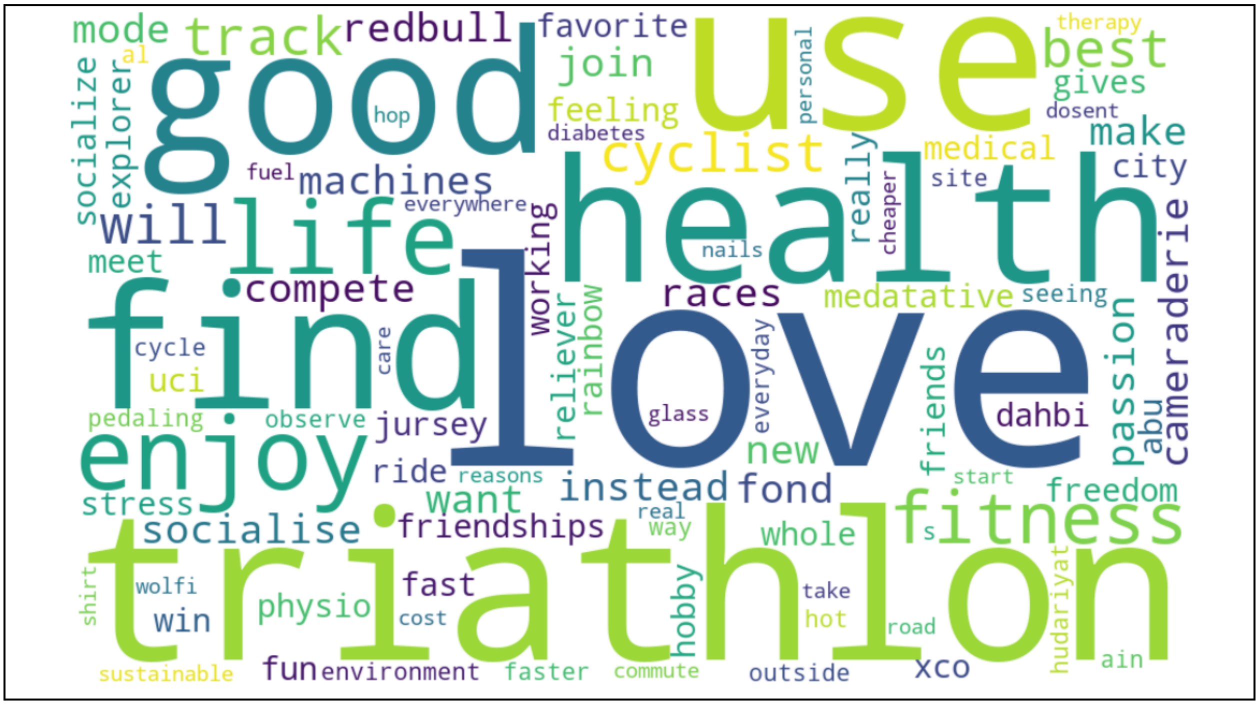


Why Ride

But it’s not this plot, it’s the word cloud that comes from the ‘other’ entries. You see bikes bring us together in a way nothing else does. It allows for the lowcost, safe, environmentally friendly, and social transportation of individuals or goods. It has no regard for wealth or class as you don’t need a special license or a large bank account to participate. Bikes (electric or not) provide a connection to your community that can be hard to find elsewhere.

I’ve learned so much about what people think when it comes to micromobility in the UAE, and I’ve been searching for exactly what gets someone out riding. With evidence supporting community building, biking events, physical/mental health, and social interactions the reason to ride became simple.

You can justify all the data you want with p-values, linear models, and random forests, but sometimes the answer is right in front of you…



“Other” reasons to ride