MGT 6203 Group Project Proposal

TEAM INFORMATION (1 point)

Team #: 83

Team Members:

1. **Amynah Reimoo, *areimoo3*:** I am a full-time student in the OMSA program. I also work part-time as a Business Intelligence Associate. My job primarily involves creating visualizations and business reports for various stakeholders. I also have a degree in Education and am a former elementary school teacher.

2. **Sidarth Sudhakar, *ssudhakar37*:** I am a part-time OMSA student working full time as a Data Scientist building enterprise grade data products. I hold a Bachelor's degree in Mechanical Engineering and have prior analytics experience predominantly in Predictive Maintenance and Supply Chain Management.

3. **Raashid Salih, *rsalih3*:** A part time OMSA student, and a recent CS graduate specializing in Data Science and Computational Intelligence. Possesses analytics experience from 8 months spent with Schlumberger in their Data and Analytics department, where I worked on tasks involving data engineering, visualization, and machine learning.

4. **Xinxing Ren, *xren81***: I am currently a student in the OMSA program while also working as an executive assistant. Previously, I held roles as a financial analyst and business intelligence analyst, where I provided strategic insights to the executive team. My educational background includes a degree in business studies.

5. **Andrew Ramirez, *aramirez89***: I am a part-time OMSA student, and recent graduate with a degree in Mathematics. I currently work as an Actuarial Analyst, and hope to transition to a more data focused job in the future.

OBJECTIVE/PROBLEM (5 points)

Project Title: Optimizing Bike Sharing Sustainability: Demand Forecasting and Inventory Management

Background Information on chosen project topic:

Bikesharing is an eco-friendly and affordable transportation choice in many urban cities across North America. Bikes are available for short term rentals across various stations in the city and stations are usually located in highly populated areas or near transit hubs. Bikes are accessible 24/7 for users and are charged based primarily on usage time and the type of bike used, which are further dependent on the users' memberships which are specific to the bike sharing company.

Capital Bikeshare is a popular bike sharing company that services the Washington DC metropolitan area in the USA with over 600 stations and 5000 bikes across 7 jurisdictions. This Bikeshare program has allowed locals and visitors to have a way to enjoy sustainable transportation options to explore the Capital of the US and generally enhances accessibility across the board.

It is important to understand that the landscape which bike sharing as a concept exists within is plagued by a certain degree of uncertainty. Customer demand can fluctuate between different stations and at different times with little concern for consistency. Hence, proper inventory management is essential to stock stations with the right amount of bikes. Understocking can lead to customer dissatisfaction and lost opportunity costs, while overstocking can lead to increased space requirements and maintenance costs. There are also costs incurred in rebalancing bikes between stations such as transportation and labor which also affects traffic and the environment. This highlights the necessity of an efficient inventory management mechanism based on robust demand forecasting (or as robust as one can get given the circumstances).

Recently, due to the Covid pandemic, Capital Bikeshare experienced a significant loss in both total cost recovery and operating cost recovery, which were down by more than 30% in 2021. Moreover, the annual membership subscription also declined by 24% compared to the previous year owing to the fact that less people were commuting and traveling during the pandemic. Consequently, it is crucial that Capital Bike Share tackles inventory management to keep the business afloat.

This problem is not only important to business owners, such as contractors for these bikes but also to the municipal government. Understanding trends in the demand, supply, cost and other ridership details can allow for smoother traffic flow, less carbon emissions and optimum profits.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

Bike Sharing operates in a very dynamic environment, and without a mechanism in place to accurately gauge demand, it may lead to losses to the company incurred from overstocking, understocking, rebalancing of resources and suboptimal pricing. This has further consequences downstream in terms of customer satisfaction, customer retention, and the environment which greatly impacts the sustainability of such a business.

State your Primary Research Question (RQ):

Which elements play a significant role in affecting demand and how can we predict the demand as correctly as possible based on this information?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Which factors have an influence on features of interest like bike share membership and ride duration? Can we quantify the significance of variables present in the original dataset?
2. How important are external factors (with respect to the dataset) like weather conditions and events in determining key variables like membership and ride duration?

3. Are there certain insights or patterns between the features that are actionable?
4. Can a sufficiently performant demand forecasting model be developed aided by said insights?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Capital bikeshare is primarily owned by member jurisdictions and operated by a contractor. The two parties may collaboratively work towards effective demand forecasting to solve the following problems:

1. Analyzing the demand and supply of bike rentals will help decision makers ensure effective inventory management

2. Special focus on the geographical data can ensure better allocation of resources, ultimately resulting in higher revenue and customer retention

3. Understanding patterns and trends in the data can help the operational teams create targeted and cost-effective marketing campaigns and optimize prices.

4. All in all, this endeavor aims to lower costs and increase revenue by making the operations more efficient.

5. The bikeshare initiative may be a part of making affordable and sustainable transportation systems available to people. Accurate demand forecasting can help governmental bodies achieve this goal.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

1. Bike share data from Capital BikeShare: https://capitalbikeshare.com/system-data
2. Weather data: https://open-meteo.com/
3. Event data: https://pypi.org/project/holidays/
4. Station data:https://opendata.dc.gov/datasets/a1f7acf65795451d89f0a38565a975b3_5/explore?filters=eyJMQVNUX1JFUE9SVEVEIjpbMTY5MDc4NjgwMDAwMCwxNjkwNzg2ODAwMDAwXX0%3D

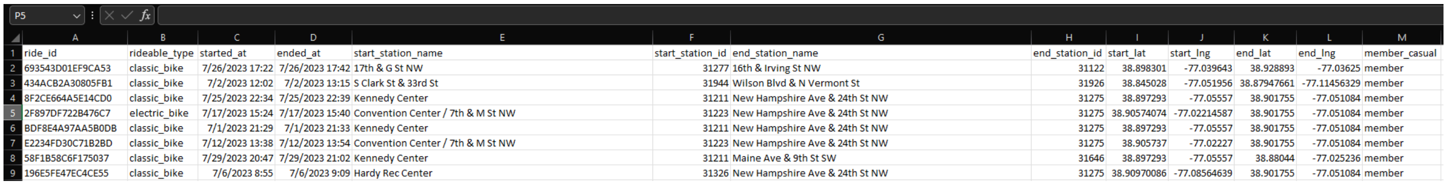
Data Description (describe each of your data sources, include screenshots of a few rows of data):

Bike Share Data

Capital Bike Share generously provides first party data based on their Data License Agreement.

- Start Date – Includes start date and time
- End Date – Includes end date and time

- Start Station – Includes starting station name and number
- End Station – Includes ending station name and number
- Bike Number – Includes ID number of bike used for the trip
- Member Type – Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)
- This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our “test” stations at our warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it's secure).



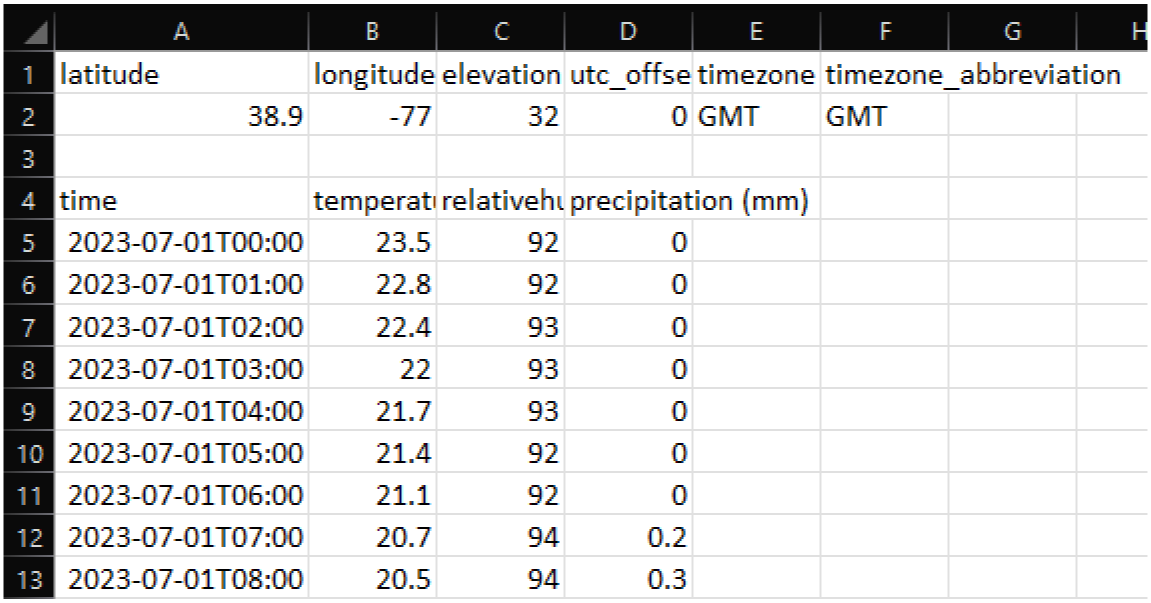
P5

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
2	693543D01EF9CA53	classic_bike	7/26/2023 17:22	7/26/2023 17:42	17th & G St NW	31277	16th & Irving St NW	31122	38.898301	-77.039643	38.928893	-77.03625	member
3	434ACB2A30805FB1	classic_bike	7/2/2023 12:02	7/2/2023 13:15	S Clark St & 33rd St	31944	Wilson Blvd & N Vermont St	31926	38.845028	-77.051956	38.87947661	-77.11456329	member
4	8F2CE664A5E14CD0	classic_bike	7/25/2023 22:34	7/25/2023 22:39	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member
5	2F897DF722B476C7	electric_bike	7/17/2023 15:24	7/17/2023 15:40	Convention Center / 7th & M St NW	31223	New Hampshire Ave & 24th St NW	31275	38.90574074	-77.02214587	38.901755	-77.051084	member
6	BDF8E4A97AA5B0DB	classic_bike	7/1/2023 21:29	7/1/2023 21:33	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member
7	E2234FD30C71B2BD	classic_bike	7/12/2023 13:38	7/12/2023 13:54	Convention Center / 7th & M St NW	31223	New Hampshire Ave & 24th St NW	31275	38.905737	-77.02227	38.901755	-77.051084	member
8	58F1B58C6F175037	classic_bike	7/29/2023 20:47	7/29/2023 21:02	Kennedy Center	31211	Maine Ave & 9th St SW	31646	38.897293	-77.05557	38.88044	-77.025236	member
9	196E5FE47EC4CE55	classic_bike	7/6/2023 8:55	7/6/2023 9:09	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.90970086	-77.08564639	38.901755	-77.051084	member

Sample Bike Share data: Includes start and end time, along with starting location and ending location for each bike used.

Weather Data

Weather API data is used to fetch different weather phenomena like temperature, humidity, wind speed both in history and 14 days into the future. The API gives us a bit of control in how to retrieve the data, by allowing us to specify the location with longitude and latitude (information we have in the Bike Share data), along with how the information is presented (for example, representing temperature in either F or C, and choosing the timezone used for the information, which will help in properly matching datasets together):



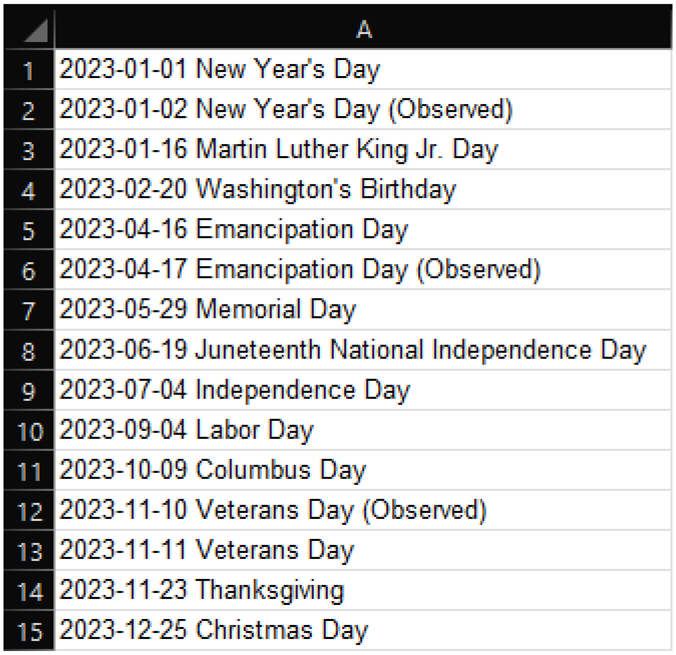
	A	B	C	D	E	F	G	H
1	latitude	longitude	elevation	utc_offse	timezone	timezone_abbreviation		
2	38.9	-77	32	0	GMT	GMT		
3								
4	time	temperat	relativeh	precipitation (mm)				
5	2023-07-01T00:00	23.5	92	0				
6	2023-07-01T01:00	22.8	92	0				
7	2023-07-01T02:00	22.4	93	0				
8	2023-07-01T03:00	22	93	0				
9	2023-07-01T04:00	21.7	93	0				
10	2023-07-01T05:00	21.4	92	0				
11	2023-07-01T06:00	21.1	92	0				
12	2023-07-01T07:00	20.7	94	0.2				
13	2023-07-01T08:00	20.5	94	0.3				

Sample weather data retrieved through API. We can specify a location, and retrieve various bits of info about the weather. Shown are the tentative variables we will be observing

Event Data

Event data source helps list recognized holidays within the country, and state. This is a bit more straightforward, and is intended to be represented as a boolean, where a date would return ‘True’ if it is a holiday, as that may impact the way in which people travel.

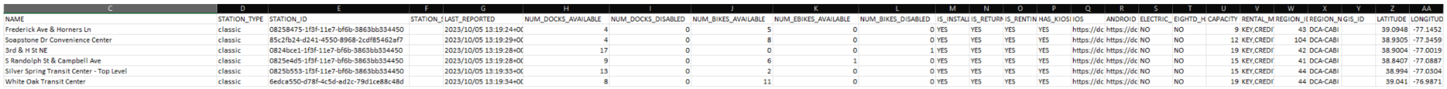
Sample:



	A
1	2023-01-01 New Year's Day
2	2023-01-02 New Year's Day (Observed)
3	2023-01-16 Martin Luther King Jr. Day
4	2023-02-20 Washington's Birthday
5	2023-04-16 Emancipation Day
6	2023-04-17 Emancipation Day (Observed)
7	2023-05-29 Memorial Day
8	2023-06-19 Juneteenth National Independence Day
9	2023-07-04 Independence Day
10	2023-09-04 Labor Day
11	2023-10-09 Columbus Day
12	2023-11-10 Veterans Day (Observed)
13	2023-11-11 Veterans Day
14	2023-11-23 Thanksgiving
15	2023-12-25 Christmas Day

Station Data

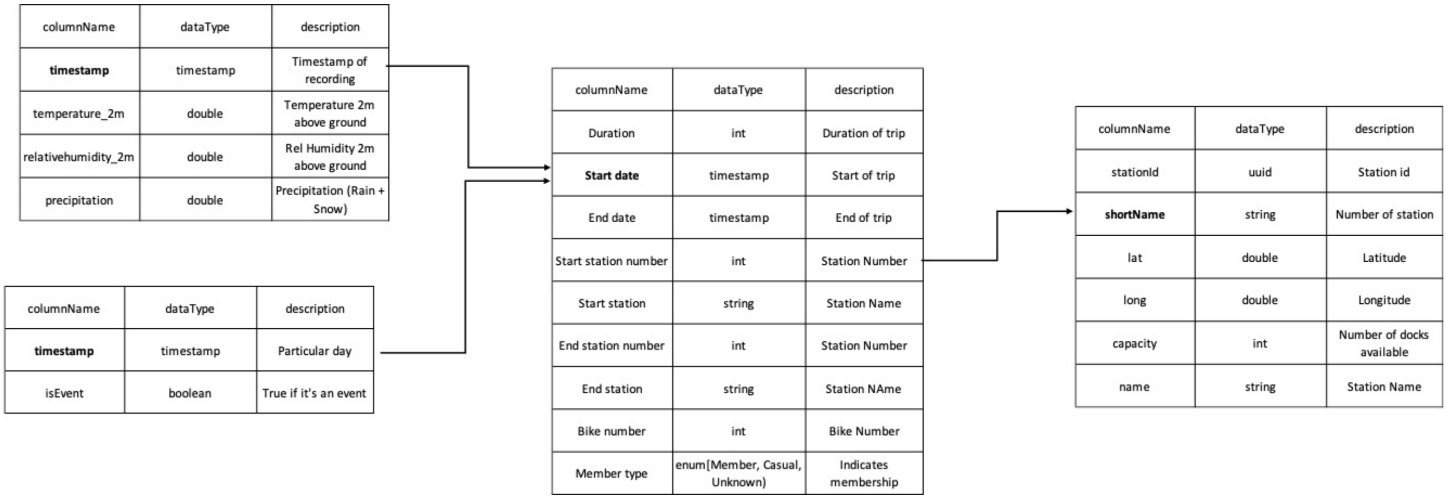
Station data is used to determine various metrics about each capital bikeshare location, such as the number of bikes available. There is also other information, such as the latitude and longitude of each location, station type (classic or lightweight), each location’s max capacity, rental methods (key or credit) and more.



C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
NAME	STATION_TYPE	STATION_ID	STATION_S	LAST_REPORTED	NUM_DOCKS_AVAILABLE	NUM_DOCKS_DISABLED	NUM_BIKES_AVAILABLE	NUM_EBIKES_AVAILABLE	NUM_BIKES_DISABLED	IS_INSTALL	IS_RETURN	IS_RENTIN	HAS_KIOSI	IOS	ANDROID	ELECTRIC_	EIGHTD_H	CAPACITY	RENTAL_M	REGION_II	REGION_N	GIS_ID	LATITUDE	LONGITUD
Frederick Ave & Horners Ln	classic	08258475-1f3f-11e7-bf6b-3863bb334450		2023/10/05 13:19:24+0C	4	0	5	0	0	YES	YES	YES	YES	https://dc	https://dc	NO	NO	9	KEY,CREDI	43	DCA-CABI		39.0948	-77.1452
Soapstone Dr Convenience Center	classic	85c2fb24-d241-4550-8968-2cdf85462af7		2023/10/05 13:19:29+0C	4	0	8	0	0	YES	YES	YES	YES	https://dc	https://dc	NO	NO	12	KEY,CREDI	104	DCA-CABI		38.9305	-77.3459
3rd & H St NE	classic	0824bce1-1f3f-11e7-bf6b-3863bb334450		2023/10/05 13:19:28+0C	17	0	0	0	1	YES	YES	YES	YES	https://dc	https://dc	NO	NO	19	KEY,CREDI	42	DCA-CABI		38.9004	-77.0019
S Randolph St & Campbell Ave	classic	0825e4d5-1f3f-11e7-bf6b-3863bb334450		2023/10/05 13:19:28+0C	9	0	6	1	0	YES	YES	YES	YES	https://dc	https://dc	NO	NO	15	KEY,CREDI	41	DCA-CABI		38.8407	-77.0887
Silver Spring Transit Center - Top Level	classic	0825b553-1f3f-11e7-bf6b-3863bb334450		2023/10/05 13:19:33+0C	13	0	2	0	0	YES	YES	YES	YES	https://dc	https://dc	NO	NO	15	KEY,CREDI	44	DCA-CABI		38.994	-77.0304
White Oak Transit Center	classic	6edca550-d78f-4c5d-ad2c-79d1ce88c48d		2023/10/05 13:19:34+0C	8	0	11	0	0	YES	YES	YES	YES	https://dc	https://dc	NO	NO	19	KEY,CREDI	44	DCA-CABI		39.041	-76.9871

Sample station data, which shows each location, along with the number of docks and bikes available at a given time

Entity Relationship Diagram



columnName	dataType	description
timestamp	timestamp	Timestamp of recording
temperature_2m	double	Temperature 2m above ground
relativehumidity_2m	double	Rel Humidity 2m above ground
precipitation	double	Precipitation (Rain + Snow)

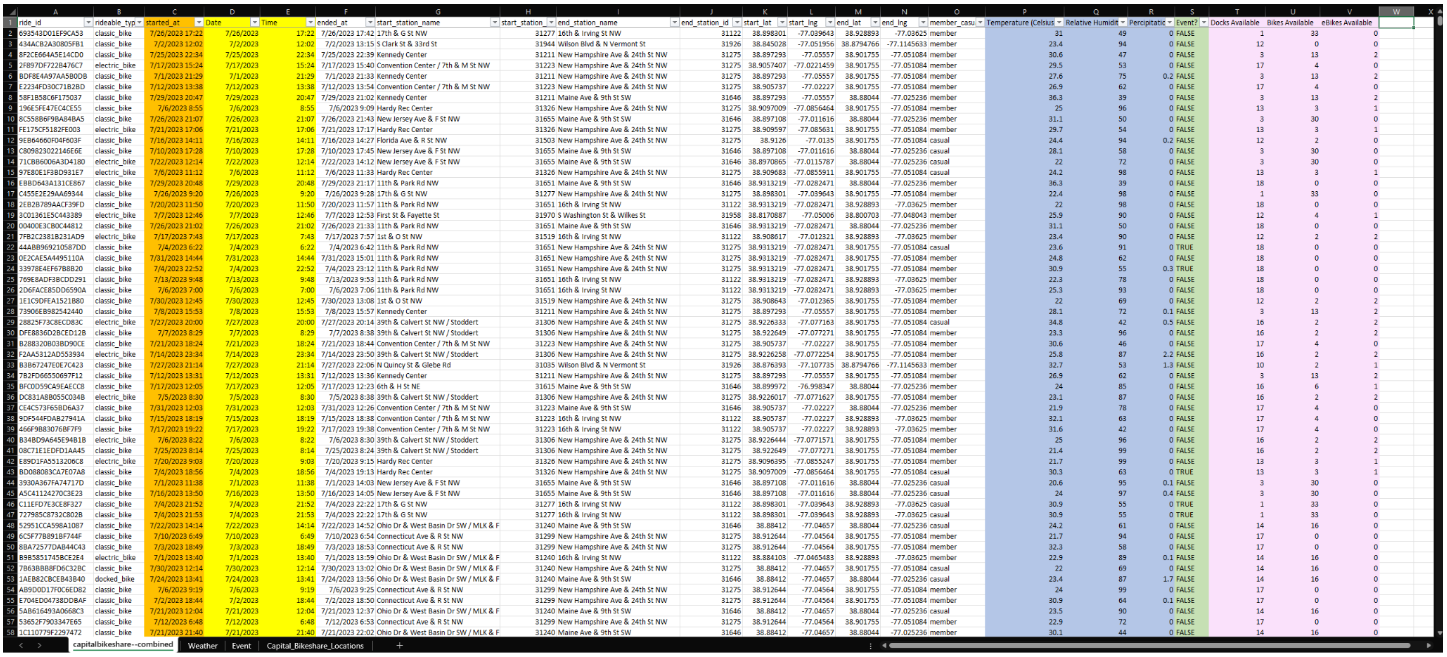
columnName	dataType	description
timestamp	timestamp	Particular day
isEvent	boolean	True if it's an event

columnName	dataType	description
Duration	int	Duration of trip
Start date	timestamp	Start of trip
End date	timestamp	End of trip
Start station number	int	Station Number
Start station	string	Station Name
End station number	int	Station Number
End station	string	Station NAme
Bike number	int	Bike Number
Member type	enum[Member, Casual, Unknown]	Indicates membership

columnName	dataType	description
stationId	uuid	Station id
shortName	string	Number of station
lat	double	Latitude
long	double	Longitude
capacity	int	Number of docks available
name	string	Station Name

The general idea of how we want to link our current information

Sample Combined data:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	ride_id	rideable_typ	started_at	Date	Time	ended_at	start_station_name	start_station_	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casu	Temperature (Celsius	Relative Humidit	Percipitatio	Event?	Docks Available	Bikes Available	eBikes Available
2	693543D01EF9CA53	classic_bike	7/26/2023 17:22	7/26/2023	17:22	7/26/2023 17:42	17th & G St NW	31277	16th & Irving St NW	31122	38.898301	-77.039643	38.928893	-77.03625	member	31	49	0	FALSE	1	33	0
3	434ACB2A30805FB1	classic_bike	7/2/2023 12:02	7/2/2023	12:02	7/2/2023 13:15	S Clark St & 33rd St	31944	Wilson Blvd & N Vermont St	31926	38.845028	-77.051956	38.8794766	-77.1145633	member	23.4	94	0	FALSE	12	0	0
4	8F2CE664A5E14CD0	classic_bike	7/25/2023 22:34	7/25/2023	22:34	7/25/2023 22:39	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member	30.6	47	0	FALSE	3	13	2
5	2F897DF722B476C7	electric_bike	7/17/2023 15:24	7/17/2023	15:24	7/17/2023 15:40	Convention Center / 7th & M St NW	31223	New Hampshire Ave & 24th St NW	31275	38.9057407	-77.0221459	38.901755	-77.051084	member	29.5	53	0	FALSE	17	4	0
6	BDF8E4A97AA5B0DB	classic_bike	7/1/2023 21:29	7/1/2023	21:29	7/1/2023 21:33	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member	27.6	75	0.2	FALSE	3	13	2
7	E2234FD30C71B2BD	classic_bike	7/12/2023 13:38	7/12/2023	13:38	7/12/2023 13:54	Convention Center / 7th & M St NW	31223	New Hampshire Ave & 24th St NW	31275	38.905737	-77.02227	38.901755	-77.051084	member	26.9	62	0	FALSE	17	4	0
8	58F1B58C6F175037	classic_bike	7/29/2023 20:47	7/29/2023	20:47	7/29/2023 21:02	Kennedy Center	31211	Maine Ave & 9th St SW	31646	38.897293	-77.05557	38.88044	-77.025236	member	36.3	39	0	FALSE	3	13	2
9	196E5FE47EC4CE55	classic_bike	7/6/2023 8:55	7/6/2023	8:55	7/6/2023 9:09	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.9097009	-77.0856464	38.901755	-77.051084	member	25	96	0	FALSE	13	3	1
10	8C558B6F9BA84BA5	classic_bike	7/26/2023 21:07	7/26/2023	21:07	7/26/2023 21:43	New Jersey Ave & F St NW	31655	Maine Ave & 9th St SW	31646	38.897108	-77.011616	38.88044	-77.025236	member	31.1	50	0	FALSE	3	30	0
11	FE175CF5182FE003	electric_bike	7/21/2023 17:06	7/21/2023	17:06	7/21/2023 17:17	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.909597	-77.085631	38.901755	-77.051084	member	29.7	54	0	FALSE	13	3	1
12	9EB64660F04F603F	classic_bike	7/16/2023 14:11	7/16/2023	14:11	7/16/2023 14:27	Florida Ave & R St NW	31503	New Hampshire Ave & 24th St NW	31275	38.9126	-77.0135	38.901755	-77.051084	casual	24.4	94	0.2	FALSE	12	2	0
13	C809823022146E6E	classic_bike	7/10/2023 17:28	7/10/2023	17:28	7/10/2023 17:45	New Jersey Ave & F St NW	31655	Maine Ave & 9th St SW	31646	38.897108	-77.011616	38.88044	-77.025236	casual	28.1	58	0	FALSE	3	30	0
14	71CBB6006A3D4180	electric_bike	7/22/2023 12:14	7/22/2023	12:14	7/22/2023 14:12	New Jersey Ave & F St NW	31655	Maine Ave & 9th St SW	31646	38.8970865	-77.0115787	38.88044	-77.025236	casual	22	72	0	FALSE	3	30	0
15	97E80E1F3BD931E7	electric_bike	7/6/2023 11:12	7/6/2023	11:12	7/6/2023 11:33	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.909683	-77.0855911	38.901755	-77.051084	casual	24.2	98	0	FALSE	13	3	1
16	EBBD643A131CE867	classic_bike	7/29/2023 20:48	7/29/2023	20:48	7/29/2023 21:17	11th & Park Rd NW	31651	Maine Ave & 9th St SW	31646	38.9313219	-77.0282471	38.88044	-77.025236	member	36.3	39	0	FALSE	18	0	0
17	C455E2E29AA69344	classic_bike	7/26/2023 9:20	7/26/2023	9:20	7/26/2023 9:28	17th & G St NW	31277	New Hampshire Ave & 24th St NW	31275	38.898301	-77.039643	38.901755	-77.051084	member	22.4	98	0	FALSE	1	33	0
18	2EB2B789AACF39FD	classic_bike	7/20/2023 11:50	7/20/2023	11:50	7/20/2023 11:57	11th & Park Rd NW	31651	16th & Irving St NW	31122	38.9313219	-77.0282471	38.928893	-77.03625	member	22	98	0	FALSE	18	0	0
19	3C01361E5C443389	electric_bike	7/7/2023 12:46	7/7/2023	12:46	7/7/2023 12:53	First St & Fayette St	31970	S Washington St & Wilkes St	31958	38.8170887	-77.05006	38.800703	-77.048043	member	25.9	90	0	FALSE	12	4	1
20	00400E3CB0C44812	classic_bike	7/26/2023 21:02	7/26/2023	21:02	7/26/2023 21:33	11th & Park Rd NW	31651	Maine Ave & 9th St SW	31646	38.9313219	-77.0282471	38.88044	-77.025236	member	31.1	50	0	FALSE	18	0	0
21	7FB2C2381B231AD9	electric_bike	7/17/2023 7:43	7/17/2023	7:43	7/17/2023 7:57	1st & O St NW	31519	16th & Irving St NW	31122	38.908617	-77.012321	38.928893	-77.03625	member	23.4	90	0	FALSE	12	2	2
22	44ABB969210587DD	classic_bike	7/4/2023 6:22	7/4/2023	6:22	7/4/2023 6:42	11th & Park Rd NW	31651	New Hampshire Ave & 24th St NW	31275	38.9313219	-77.0282471	38.901755	-77.051084	casual	23.6	91	0	TRUE	18	0	0
23	0E2CAE5A4495110A	classic_bike	7/31/2023 14:44	7/31/2023	14:44	7/31/2023 15:01	11th & Park Rd NW	31651	New Hampshire Ave & 24th St NW	31275	38.9313219	-77.0282471	38.901755	-77.051084	member	24.8	62	0	FALSE	18	0	0
24	33978E4EF67B8B20	classic_bike	7/4/2023 22:52	7/4/2023	22:52	7/4/2023 23:12	11th & Park Rd NW	31651	New Hampshire Ave & 24th St NW	31275	38.9313219	-77.0282471	38.901755	-77.051084	member	30.9	55	0.3	TRUE	18	0	0
25	769E8ADF3BCDD291	classic_bike	7/13/2023 9:48	7/13/2023	9:48	7/13/2023 9:53	11th & Park Rd NW	31651	16th & Irving St NW	31122	38.9313219	-77.0282471	38.928893	-77.03625	member	22.3	78	0	FALSE	18	0	0
26	2D6FACE85DD6590A	classic_bike	7/6/2023 7:00	7/6/2023	7:00	7/6/2023 7:06	11th & Park Rd NW	31651	16th & Irving St NW	31122	38.9313219	-77.0282471	38.928893	-77.03625	member	25.3	93	0	FALSE	18	0	0
27	1E1C9DFEA1521B80	classic_bike	7/30/2023 12:45	7/30/2023	12:45	7/30/2023 13:08	1st & O St NW	31519	New Hampshire Ave & 24th St NW	31275	38.908643	-77.012365	38.901755	-77.051084	member	22	69	0	FALSE	12	2	2
28	73906EB982542440	classic_bike	7/8/2023 15:53	7/8/2023	15:53	7/8/2023 15:57	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member	28.1	72	0.1	FALSE	3	13	2
29	28825F73C8ECD83C	electric_bike	7/27/2023 20:00	7/27/2023	20:00	7/27/2023 20:14	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.9226333	-77.077163	38.901755	-77.051084	casual	34.8	42	0.5	FALSE	16	2	2
30	DFE8836D2BCED12B	classic_bike	7/7/2023 8:29	7/7/2023	8:29	7/7/2023 8:38	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.922649	-77.077271	38.901755	-77.051084	member	23.3	96	0	FALSE	16	2	2
31	B288320B03BD90CE	classic_bike	7/21/2023 18:24	7/21/2023	18:24	7/21/2023 18:44	Convention Center / 7th & M St NW	31223	New Hampshire Ave & 24th St NW	31275	38.905737	-77.02227	38.901755	-77.051084	member	30.6	46	0	FALSE	17	4	0
32	F2AA5312AD553934	electric_bike	7/14/2023 23:34	7/14/2023	23:34	7/14/2023 23:50	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.9226258	-77.0772254	38.901755	-77.051084	member	25.8	87	2.2	FALSE	16	2	2
33	B3B67247E0E7C423	classic_bike	7/27/2023 21:14	7/27/2023	21:14	7/27/2023 22:06	N Quincy St & Glebe Rd	31035	Wilson Blvd & N Vermont St	31926	38.876393	-77.107735	38.8794766	-77.1145633	member	32.7	53	1.3	FALSE	10	2	1
34	7B2FD66550697F12	classic_bike	7/12/2023 13:31	7/12/2023	13:31	7/12/2023 13:36	Kennedy Center	31211	New Hampshire Ave & 24th St NW	31275	38.897293	-77.05557	38.901755	-77.051084	member	26.9	62	0	FALSE	3	13	2
35	BFC0D59CA9EAECC8	classic_bike	7/17/2023 12:05	7/17/2023	12:05	7/17/2023 12:23	6th & H St NE	31615	Maine Ave & 9th St SW	31646	38.899972	-76.998347	38.88044	-77.025236	member	24	85	0	FALSE	16	6	1
36	DC831A8B055C034B	electric_bike	7/5/2023 8:30	7/5/2023	8:30	7/5/2023 8:38	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.9226017	-77.0771627	38.901755	-77.051084	member	23.1	87	0	FALSE	16	2	2
37	CE4C573F65BD6A37	classic_bike	7/31/2023 12:03	7/31/2023	12:03	7/31/2023 12:26	Convention Center / 7th & M St NW	31223	Maine Ave & 9th St SW	31646	38.905737	-77.02227	38.88044	-77.025236	member	21.9	78	0	FALSE	17	4	0
38	9DF544FDAB27941A	classic_bike	7/15/2023 18:19	7/15/2023	18:19	7/15/2023 18:38	Convention Center / 7th & M St NW	31223	16th & Irving St NW	31122	38.905737	-77.02227	38.928893	-77.03625	member	32.1	63	0	FALSE	17	4	0
39	466F9B83076BF7F9	classic_bike	7/17/2023 19:22	7/17/2023	19:22	7/17/2023 19:38	Convention Center / 7th & M St NW	31223	16th & Irving St NW	31122	38.905737	-77.02227	38.928893	-77.03625	member	31.6	42	0	FALSE	17	4	0
40	B34BD9A645E94B1B	electric_bike	7/6/2023 8:22	7/6/2023	8:22	7/6/2023 8:30	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.9226444	-77.0771571	38.901755	-77.051084	member	25	96	0	FALSE	16	2	2
41	08C71E1EDFD1AA45	classic_bike	7/25/2023 8:14	7/25/2023	8:14	7/25/2023 8:24	39th & Calvert St NW / Stoddert	31306	New Hampshire Ave & 24th St NW	31275	38.922649	-77.077271	38.901755	-77.051084	member	21.4	99	0	FALSE	16	2	2
42	E89D1FA5513206C8	electric_bike	7/20/2023 9:03	7/20/2023	9:03	7/20/2023 9:15	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.9096395	-77.0855247	38.901755	-77.051084	member	21.7	99	0	FALSE	13	3	1
43	BD088083CA7E07A8	classic_bike	7/4/2023 18:56	7/4/2023	18:56	7/4/2023 19:13	Hardy Rec Center	31326	New Hampshire Ave & 24th St NW	31275	38.9097009	-77.0856464	38.901755	-77.051084	casual	30.3	63	0	TRUE	13	3	1
44	3930A367FA74717D	classic_bike	7/1/2023 11:38	7/1/2023	11:38	7/1/2023 14:03	New Jersey Ave & F St NW	31655	Maine Ave & 9th St SW	31646	38.897108	-77.011616	38.88044	-77.025236	casual	20.6	95	0.1	FALSE	3	30	0
45	A5C41124270C3E23	classic_bike	7/16/2023 13:50	7/16/2023	13:50	7/16/2023 14:05	New Jersey Ave & F St NW	31655	Maine Ave & 9th St SW	31646	38.897108	-77.011616	38.88044	-77.025236	casual	24	97	0.4	FALSE	3	30	0
46	C11EFD7E3CE8F327	classic_bike	7/4/2023 21:52	7/4/2023	21:52	7/4/2023 22:22	17th & G St NW	31277	16th & Irving St NW	31122	38.898301	-77.039643	38.928893	-77.03625	casual	30.9	55	0	TRUE	1	33	0
47	727985C8732C802B	classic_bike	7/4/2023 21:53	7/4/2023	21:53	7/4/2023 22:22	17th & G St NW	31277	16th & Irving St NW	31122	38.898301	-77.039643	38.928893	-77.03625	casual	30.9	55	0	TRUE	1	33	0
48	52951CCA598A1087	classic_bike	7/22/2023 14:14	7/22/2023	14:14	7/22/2023 14:52	Ohio Dr & West Basin Dr SW / MLK & F	31240	Maine Ave & 9th St SW	31646	38.88412	-77.04657	38.88044	-77.025236	casual	24.2	61	0	FALSE	14	16	0
49	6C5F77B891BF744F	classic_bike	7/10/2023 6:49	7/10/2023	6:49	7/10/2023 6:54	Connecticut Ave & R St NW	31299	New Hampshire Ave & 24th St NW	31275	38.912644	-77.04564	38.901755	-77.051084	member	21.7	94	0	FALSE	17	0	0
50	8BA72577DAB44C43	classic_bike	7/3/2023 18:49	7/3/2023	18:49	7/3/2023 18:53	Connecticut Ave & R St NW	31299	New Hampshire Ave & 24th St NW	31275	38.912644	-77.04564	38.901755	-77.051084	member	32.3	58	0	FALSE	17	0	0
51	B9B5851745BCE2E4	electric_bike	7/1/2023 13:40	7/1/2023	13:40	7/1/2023 13:59	Ohio Dr & West Basin Dr SW / MLK & F	31240	16th & Irving St NW	31122	38.884103	-77.0465483	38.928893	-77.03625	member	22.9	89	0.1	FALSE	14	16	0
52	7B63BBB8FD6C32BC	classic_bike	7/30/2023 12:14	7/30/2023	12:14	7/30/2023 13:02	Ohio Dr & West Basin Dr SW / MLK & F	31240	New Hampshire Ave & 24th St NW	31275	38.88412	-77.04657	38.901755	-77.051084	casual	22	69	0	FALSE	14	16	0
53	1AEB82CBCEB43B40	docked_bike	7/24/2023 13:41	7/24/2023	13:41	7/24/2023 13:56	Ohio Dr & West Basin Dr SW / MLK & F	31240	Maine Ave & 9th St SW	31646	38.88412	-77.04657	38.88044	-77.025236	casual	23.4	87	1.7	FALSE	14	16	0
54	AB9D0D17F0C6ED82	classic_bike	7/6/2023 9:19	7/6/2023	9:19	7/6/2023 9:25	Connecticut Ave & R St NW	31299	New Hampshire Ave & 24th St NW	31275	38.912644	-77.04564	38.901755	-77.051084	member	24	99	0	FALSE	17	0	0
55	E704ED04738DDBAF	classic_bike	7/2/2023 18:44	7/2/2023	18:44	7/2/2023 18:50	Connecticut Ave & R St NW	31299	New Hampshire Ave & 24th St NW	31275	38.912644	-77.04564	38.901755	-77.051084	member	30.9	64	0.1	FALSE	17	0	0
56	5AB616493A0668C3	classic_bike	7/21/2023 12:04	7/21/2023	12:04	7/21/2023 12:37	Ohio Dr & West Basin Dr SW / MLK & F	31240	Maine Ave & 9th St SW	31646	38.88412	-77.04657	38.88044	-77.025236	casual	23.5	90	0	FALSE	14	16	0
57	53652F7903347E65	classic_bike	7/12/2023 6:48	7/12/2023	6:48	7/12/2023 6:53	Connecticut Ave & R St NW	31299	New Hampshire Ave & 24th St NW	31275	38.912644	-77.04564	38.901755	-77.051084	member	22.9	72	0	FALSE	17	0	0
58	1C110779F2297472	classic_bike	7/21/2023 21:40	7/21/2023	21:40	7/21/2023 22:02	Ohio Dr & West Basin Dr SW / MLK & F	31240	Maine Ave & 9th St SW	31646	38.88412	-77.04657	38.88044	-77.025236	member	30.1	44	0	FALSE	14	16	0

An example of how the data would look when combined, color-coded to show how different sources are merged

Grouped together as described in the ERD: The orange column (date + time) works as a primary key to get data from the weather (blue columns), and event (green column) data sets. The station name is used to get the info from the station (pink columns) data sets. The yellow columns are just to split the date and time into their own column, which may be helpful in portraying, for example, seasonal trends within the data.

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Exploratory Data Analysis

Dependent variables are going to be ride duration, ride type and user membership. These are key in our analysis as they are correlated with profit.

Independent variables are station information and might include the above in case of univariate analysis.

Ride duration is to be derived from start and end time.

Weather information constitutes an independent variable. However, new features need to be engineered that boil down different metrics (temperature, precipitation, humidity) into clean bins like (sunny, cloudy, rainy).

Event data is also an independent variable. It likewise has to be engineered in a way that delineates any particular day as the weekend, public holiday, or otherwise.

Demand Forecasting

For demand forecasting, potential independent variables are going to be temporal variables which are engineered from the ride dataset. These include:

- Time of the day
- Day of the week
- Season/Month of the year - Season to be encoded based on the month.

Similarly, weather information could be aggregated using their average data during these engineered temporal variables. From the Weather API, key variables include:

- Humidity
- Temperature
- Precipitation

Additionally, to understand the impact of any event (holidays), a boolean variable has to be created.

The demand, which is the dependent variable, is going to be the aggregated count of rides grouped on all the temporal independent variables.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Data Preparation

1. Obtain a subset of data from the source since we have data spanning the years 2010 to July 2023.
    a. Using all of the data can be a poor decision because less recent data will not be relevant to more recent data, hence increasing noise.
    b. On the other hand, just utilizing the most recent data will decrease the amount of data on hand, yielding results that are not generalizable.
    c. Furthermore, we might omit data pertaining to right after CoVID began since the ramifications of the era might increase noise as well.
    d. Thus, to strike a balance between these different considerations, we have ***tentatively*** opted to work on data from 2015 up until 2020.
2. Obtain external data for weather and events, either from an API or using a dataset.
3. Merge the data together to create the master dataset.
4. Implement data cleaning methods on said dataset. This includes, but is not limited to:

   a. Dealing with missing data
   b. Removing potential duplicates
   c. Outlier identification
   d. Depending on the results, outlier removal
   e. Ensuring appropriate data types
5. Perform feature engineering to obtain certain derived columns of interest. (Split date into day, month, etc.)
6. Finally, make the dataset available for Exploratory Data Analysis and Demand Forecasting Modeling.

Exploratory Data Analysis

1. We begin by carefully inspecting the data types and the structure.
2. We can take a look at the statistical summary of the dataset to further understand the mean, median and standard deviation, etc. (especially for numerical features)
3. We will then conduct univariate and multivariate analysis to answer certain hypotheses and gain useful insights. This is fairly important for specifically gauging the relationships between the dependent and independent variables.
4. We will also model certain features of interest to quantify the degree of correlation and also the significance.
5. Explainable models like linear regression and decision trees will be employed in this regard to obtain actionable information.
6. Certain transformations (like log transformation, for instance) might be applied to make the data more conducive to modeling efforts.
7. Models will be compared using various metrics to get a nuanced understanding of fit, but generally adjusted R^2 is what is going to be used to facilitate the comparison.

Demand Forecasting

1. Analyzing the impact of each engineered feature by visualizing the distribution of demand across each of them.
2. Grouping these features into well-defined clusters to reduce cardinality in these categorical variables
3. Understanding if there are any inherent clusters with each docking station's geographical location.
   a. Station level information might suffer from high cardinality. Whereas clusters are more intuitive. Riders would also be motivated to take a bike from nearby docking stations.
4. Can potentially use log transformation to remove the skewness in the observed distributions.
5. Two potential modeling techniques
   a. Demand Prediction using a tree based / linear model depending on the distribution and use forecasted weather information to predict the demand for next n days
   b. Build a forecasting technique using ARIMA / Bayesian Regression just using the underlying time Series data [Without Weather] and forecast the demand with a given confidence.
6. Similarly modeling could be done separately for registered riders and casual riders.
7. Model Comparison should be done solely based on the percentage error on test data like MAPE for example to quantify the accuracy of predicting next n observations.
8. Hyperparameter tuning can be considered depending on the model used.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

1. Hourly trend in demand as users generally use bikes for their daily commute either to office/college.
2. Demand potentially would increase during the weekdays over weekends.
3. Rain would be the highest contributing factor among other weather phenomena.
4. Certain neighborhoods/locations would face higher demand due to close proximity to prominent destinations [Office/School/Market].
5. Rides in general, and ride duration probably takes a dip during bad weather.
6. Poor weather might affect casual riders more than members.
7. Certain stations might be overrepresented during certain events.
8. Likewise, demand for certain stations drops during public holidays and events since there are no people heading to work or school.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

1. Location of bike stations and the number of bikes in each station would be adjusted based on demand forecasting and inventory optimisation.
2. The insights gained from the analysis would provide management with a better idea of the various parameters that drive key interests, and to customer behavior. This can lead to more informed decision making throughout.
3. The offers that are being rolled out to customers to keep them engaged will be tweaked to prevent customer churn and incentivize bike usage during low demand season.
4. Ultimately, all these decisions would increase the bottom-line revenue for the company whilst enhancing customers engagement.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Stage	Task	Description	Deadline
Proposal	Proposal Document	Prepare Proposal Document	8-Oct-23
Data Preparation	Obtain a subset of data from the source	Filter the data from 2015 to 2020, excluding the CoVID period	22-Oct-23
	Obtain external data for weather and	Use an API or a dataset to get the relevant	

	events	information	
	Merge the data together to create the master dataset	Join the data on common keys and check for consistency	
	Implement data cleaning methods on the dataset	Handle missing data, duplicates, outliers, and data types	
	Perform feature engineering to obtain derived columns	Split date into day, month, etc. and create other features of interest	
Progress Update	Progress Report	Prepare Progress Report	5-Nov-23
Exploratory Data Analysis	Inspect the data types and structure	Use descriptive statistics and summary functions to check the data quality and format	27-Nov-23
	Conduct univariate and multivariate analysis	Use plots and tables to visualize the distribution and relationship of the variables	
	Derive insights from basic analysis	Use results from plots and tables to obtain actionable insights	
	Apply certain transformations to the data	Use log transformation or other methods to make the data more suitable for modeling	
	Test certain hypotheses and model certain relationships	Use statistical tests and explainable models to quantify and validate the findings	
	Determine viable results	Narrow down on insights of statistical significance	
Demand Forecasting	Analyze the impact of each engineered feature on demand	Use plots and tables to visualize the demand across each feature and cluster	27-Nov-23
	Build different models to predict demand for bike rentals	Use tree based or linear models with weather information or time series models without weather information	
	Analyze models for possible insights	Use model parameters to gauge seasonality of certain variables and other potentially useful information	
	Evaluate and compare different models based on percentage error on test data	Use MAPE or other metrics to measure the accuracy of predicting next n observations	
	Model separately for registered riders and casual riders	Repeat the above steps for each segment of customers and compare the results	

Final Report	Final Report	Prepare Final Report	4-Dec-23