MGT 6203 Group Project Proposal

TEAM INFORMATION: Team #104

Team Members:

1. **Caroline Schmitt** (GTID: cschmitt33) - Caroline received her BS in Statistics and BA in German from the University of Georgia. She is currently enrolled in the Georgia Tech OMSA program.
2. **Emy Ng** (GTID: ehang6) – Emy is a Data Analyst in the banking industry specifically in the commercial department with a BS in Computer Information System from University of Louisville in 2014.
3. **Matthew Kim** (GTID: mkim629) - Matt received his BBA in Finance from the University of Georgia. He currently works at the Federal Reserve Bank of Atlanta as a Data Analyst.
4. **Mike Genovese** (GTID: mgenovese6) - Mike currently works as Tech Consultant who graduated from Virginia Tech with a BS in Business Information Technology. He began his graduate work in the Georgia Tech OMSA program in 2022.
5. **Osman Yardimci** (GTID: oyardimci3) - Osman is a faculty member at Tuskegee University, Alabama. He received his Ph.D. from Auburn University in 2019.

OBJECTIVE/PROBLEM

Project Title: How Do US Gas Prices Affect Alternative Methods of Travel?

Background Information on chosen project topic: The United States has been a historically car dependent society. But with fluctuating gas prices, alternative methods of travel such as public transportation, carpooling, and biking are viable options to avoid spending more at the pump.

Methods of transportation are substitutable from each other. An individual can choose to travel to a destination by car, by rail, or by bike - but may only select a single mode at a time. Therefore, it may be useful to examine how the economics of refueling automobiles can affect consumer behavior, and ultimately what alternates they may (or may not) choose. Furthermore, it may be interesting to explore these choices against a quantifiable measure such as air quality. Since not all forms of transportation are created equal (i.e. an automobile with a single rider will create much more pollution per capita than a subway), a statistically significant observation could be obtained by comparing all of these datasets.

Problem Statement: Alternative methods of transportation like public transit and bike share programs are affordable options for people to avoid spending on gasoline for personal automobiles, but ridership trends for these alternative methods aren’t publicly modeled. The purpose of our analysis will be to provide a model for transportation organizations based on monthly US gas prices.

Primary Research Question (RQ): How does the average US gas price over time affect ridership numbers of alternative transportation methods?

Supporting Research Questions:

1. Which areas show the most increase in transit ridership in response to increased gas prices?
2. Do all grades of gasoline fluctuate in price at roughly the same rate? And if not, does one affect alternative transportation ridership more than others?
3. Are there any particular significant events that explain any sudden spikes in alternative transportation ridership numbers?

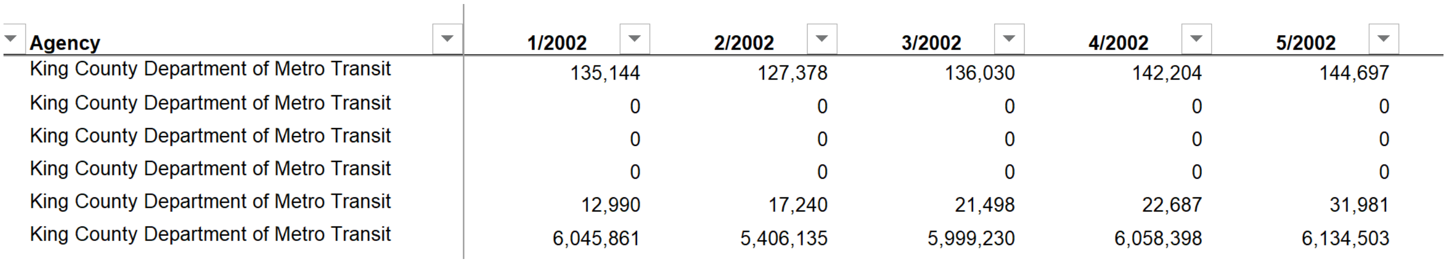
Business Justification: This research question can help an alternative transportation organization or company model demand if gas prices are correlated to ridership. Organizations won’t miss out on potential profits by having insufficient resources to meet demand and won’t have a surplus of resources when ridership is lower.

DATASET SOURCES AND DESCRIPTIONS

Data Sources:

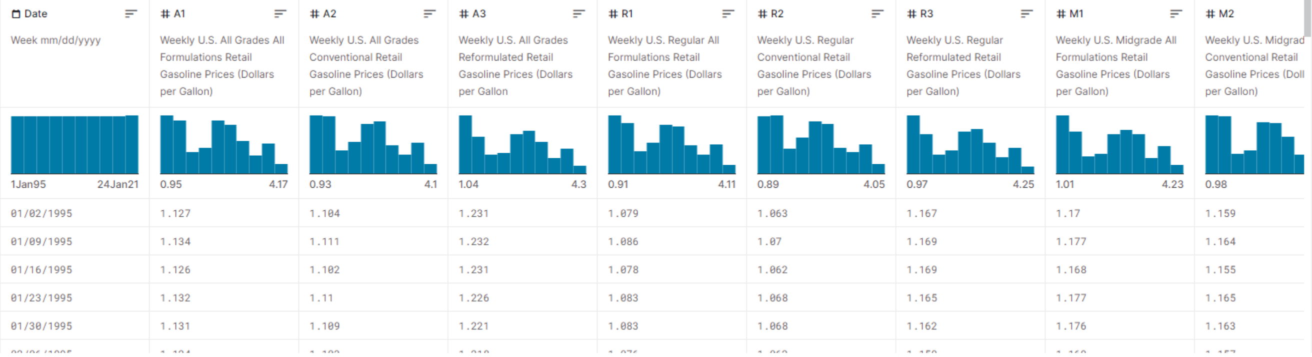
- Complete Monthly Ridership (with adjustments and estimates)
- U.S. Gasoline and Diesel Retail Prices 1995-2021
- US Air Quality 1980-Present
- Weekly Petroleum Status Report (possible use -- see appendix for description)
- Bike Sharing Dataset (possible use -- see appendix for description)

Monthly ridership: This is a time-series dataset spanning from 1/2002 to 4/2023. There are approximately 2,200 rows, one per regional transit provider and mode (note that below, modality details are concealed.)



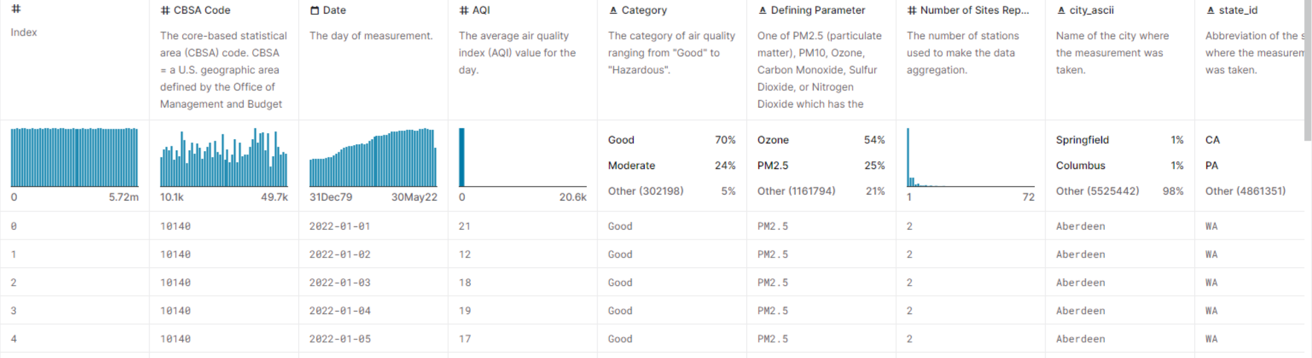
Agency	1/2002	2/2002	3/2002	4/2002	5/2002
King County Department of Metro Transit	135,144	127,378	136,030	142,204	144,697
King County Department of Metro Transit	0	0	0	0	0
King County Department of Metro Transit	0	0	0	0	0
King County Department of Metro Transit	0	0	0	0	0
King County Department of Metro Transit	12,990	17,240	21,498	22,687	31,981
King County Department of Metro Transit	6,045,861	5,406,135	5,999,230	6,058,398	6,134,503

U.S. Gasoline and Diesel Retail Prices 1995-2021: Time-series data that shows weekly retail prices in the United States for multiple formulations of gasoline, 1995 to 2021.



Date	A1	A2	A3	R1	R2	R3	M1	M2
Week mm/dd/yyyy	Weekly U.S. All Grades All Formulations Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. All Grades Conventional Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. All Grades Reformulated Retail Gasoline Prices (Dollars per Gallon	Weekly U.S. Regular All Formulations Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. Regular Conventional Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. Regular Reformulated Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. Midgrade All Formulations Retail Gasoline Prices (Dollars per Gallon)	Weekly U.S. Midgrad Conventional Retail Gasoline Prices (Doll per Gallon)
1Jan95 24Jan21	0.95 4.17	0.93 4.1	1.04 4.3	0.91 4.11	0.89 4.05	0.97 4.25	1.01 4.23	0.98
01/02/1995	1.127	1.104	1.231	1.079	1.063	1.167	1.17	1.159
01/09/1995	1.134	1.111	1.232	1.086	1.07	1.169	1.177	1.164
01/16/1995	1.126	1.102	1.231	1.078	1.062	1.169	1.168	1.155
01/23/1995	1.132	1.11	1.226	1.083	1.068	1.165	1.177	1.165
01/30/1995	1.131	1.109	1.221	1.083	1.068	1.162	1.176	1.163

US Air Quality 1980-Present: Time-series data that shows daily air quality index values and other categorical values to measure air quality by region in the United States.



#	CBSA Code	Date	AQI	Category	Defining Parameter	Number of Sites Rep...	city_ascii	state_id
Index	The core-based statistical area (CBSA) code. CBSA = a U.S. geographic area defined by the Office of Management and Budget	The day of measurement.	The average air quality index (AQI) value for the day.	The category of air quality ranging from "Good" to "Hazardous".	One of PM2.5 (particulate matter), PM10, Ozone, Carbon Monoxide, Sulfur Dioxide, or Nitrogen Dioxide which has the	The number of stations used to make the data aggregation.	Name of the city where the measurement was taken.	Abbreviation of the s where the measurem was taken.
0 5.72m	10.1k 49.7k	31Dec79 30May22	0 20.6k	Good 70% Moderate 24% Other (302198) 5%	Ozone 54% PM2.5 25% Other (1161794) 21%	1 72	Springfield 1% Columbus 1% Other (5525442) 98%	CA PA Other (4861351)
0	10140	2022-01-01	21	Good	PM2.5	2	Aberdeen	WA
1	10140	2022-01-02	12	Good	PM2.5	2	Aberdeen	WA
2	10140	2022-01-03	18	Good	PM2.5	2	Aberdeen	WA
3	10140	2022-01-04	19	Good	PM2.5	2	Aberdeen	WA
4	10140	2022-01-05	17	Good	PM2.5	2	Aberdeen	WA

Key Variables:

Dependent: The monthly ridership variable within each alternative transportation dataset.

Independent: Calculated average monthly US gas prices for all grades, including regular grade, midgrade, premium grade, and diesel. We hypothesize that regular grade will be the most important independent variable if the four different grades aren't correlated and can all be used in the model.

APPROACH/METHODOLOGY

Planned Approach: We plan to try a variety of time series models, mostly multivariate models, including vector autoregressive (VAR) models; generalized autoregressive conditional heteroskedasticity (GARCH) models; decomposition models; and others. We will explore multiple approaches for hyperparameter optimization including exploratory analysis such as analyzing ACF and PACF plots to inform VAR order in addition to computational strategies like grid searching and auto-ARIMA. To model, we will likely aggregate data to different time periods. Some models may require transforming data to be centered or stationary or for zeros to be transformed. For price data, we may need to adjust for inflation.

For this project, inference is more important than prediction. When evaluating and comparing models, we'll be concerned with model accuracy, measured by e.g. MAE, MAPE, RMSE as evaluated on a holdout set, but goodness-of-fit measures will also be critical. For inference, model interpretability matters. Thus we may prefer lower-performing but interpretable models with testable assumptions over higher-performing but black box-style models such as neural nets.

Anticipated Conclusions/Hypothesis: We're expecting some relationship between gas prices, air quality, and transit ridership. In particular we expect rises in transit ridership when gas prices are especially high. But gas prices, air quality, and transit ridership are all complicated and affected by many exogenous variables not explicitly in our dataset. We don't expect to be able to explain any of our variables of interest completely.

What business decisions will be impacted by the results of your analysis? What could be some benefits? Ridership forecasting is useful for transit agencies as well as companies like Lyft or Uber that compete with public transit. Understanding regional ridership elasticity may also help regional and national policymakers increase public transit use.

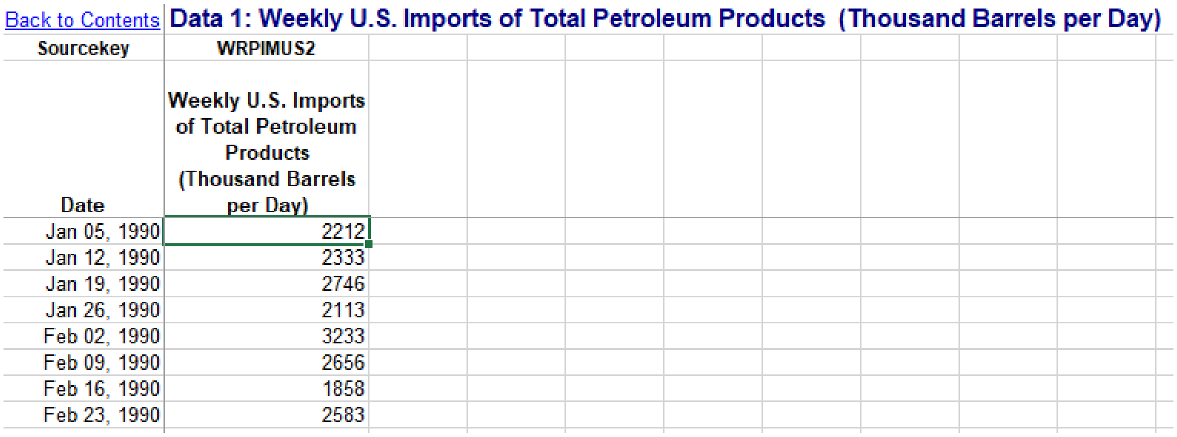
PROJECT TIMELINE/PLANNING

Project timeline/key dates: We hope to start our slides for the proposal video on the 22nd. We hope to have preliminary data cleaning and some EDA done by July 1st and some initial models fit by July 5th, in time for the July 9th progress report.

APPENDIX

Weekly Petroleum Status Report

Time-series data that shows the weekly imported oil barrels. The site also has an option for specifically just total motor gasoline or all petroleum products.

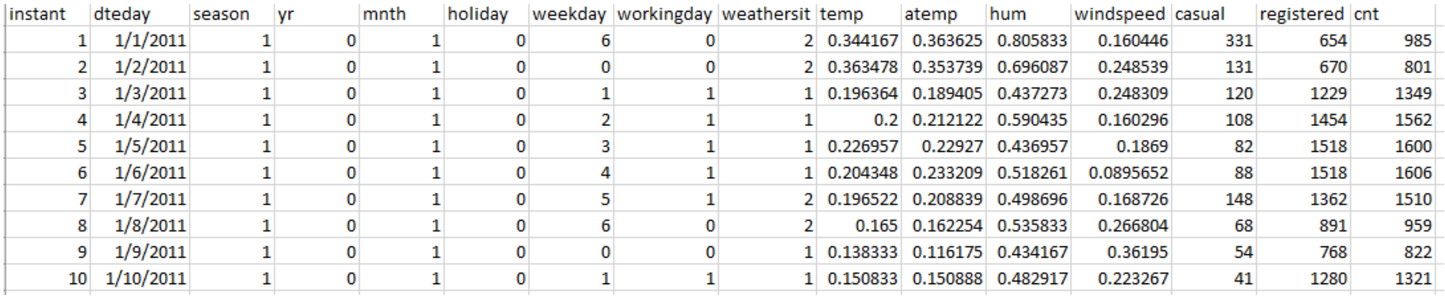


Back to Contents **Data 1: Weekly U.S. Imports of Total Petroleum Products (Thousand Barrels per Day)**

Sourcekey	WRPIMUS2
Date	Weekly U.S. Imports of Total Petroleum Products (Thousand Barrels per Day)
Jan 05, 1990	2212
Jan 12, 1990	2333
Jan 19, 1990	2746
Jan 26, 1990	2113
Feb 02, 1990	3233
Feb 09, 1990	2656
Feb 16, 1990	1858
Feb 23, 1990	2583

Bike Sharing Dataset

Time-series data for 2011 to 2012 that shows the daily Capital bike sharing ridership as well as variables such as date, time of week, and weather.



instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	1/1/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	1/2/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	1/3/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	1/4/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	1/5/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	1/6/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
7	1/7/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	1/8/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	1/9/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	1/10/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321