PROJECT PORTFOLIO

Syracuse University – School of Information Studies
2022

M.S APPLIED DATA SCIENCE

Marley Akonnor
SUID: 662128458
Email: makonnor@syr.edu
Github: https://github.com/Makonnor1/MSDS-Portfolio

Table of Contents:

Introduction:

The M.S. in Applied Data Science from Syracuse University is touted as a practitioner's degree strongly rooted in project-based learning. The curriculum deliverables for courses consisted of four primary tenets. First, data collection and organization by way of various tools. Second, data analysis through pattern recognition made available through carefully constructed visualizations, data mining, and statistical analysis. Third, demonstration of the ability to make decisions and execute strategy based on the data gathered. Finally, the ability to implement a sound and cogent plan of action to further business goals.

The emphasis of these principals ensured a repeatable mental framework for approaching all projects as well as enterprise level endeavors with the diverse data science foundation of data capture, management, analysis, and communication. The following courses and accompanying projects are put forth as an acknowledgement and confirmation of curricular outcomes solidified during my tenure at the School of Information Studies:

1) IST 659 – Database Administration Concepts & Database Management
2) IST 652 – Scripting for Data Analysis
3) IST 718 – Big Data Analytics

The stated goals of the M.S in Applied Data Science Portfolio, are for students to demonstrate to faculty the following seven learning objectives:

1. Describe a broad overview of the major practice areas in data science
2. Collect and organize data
3. Identify patterns in data via visualization, statistical analysis, and data mining
4. Develop alternative strategies based on the data
5. Develop a plan of action to implement the business decisions derived from the analyses

6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

This portfolio will confirm that.

Project 1:

IST 659: Database Administration Concepts & Database Management
Professor: Chad Harper

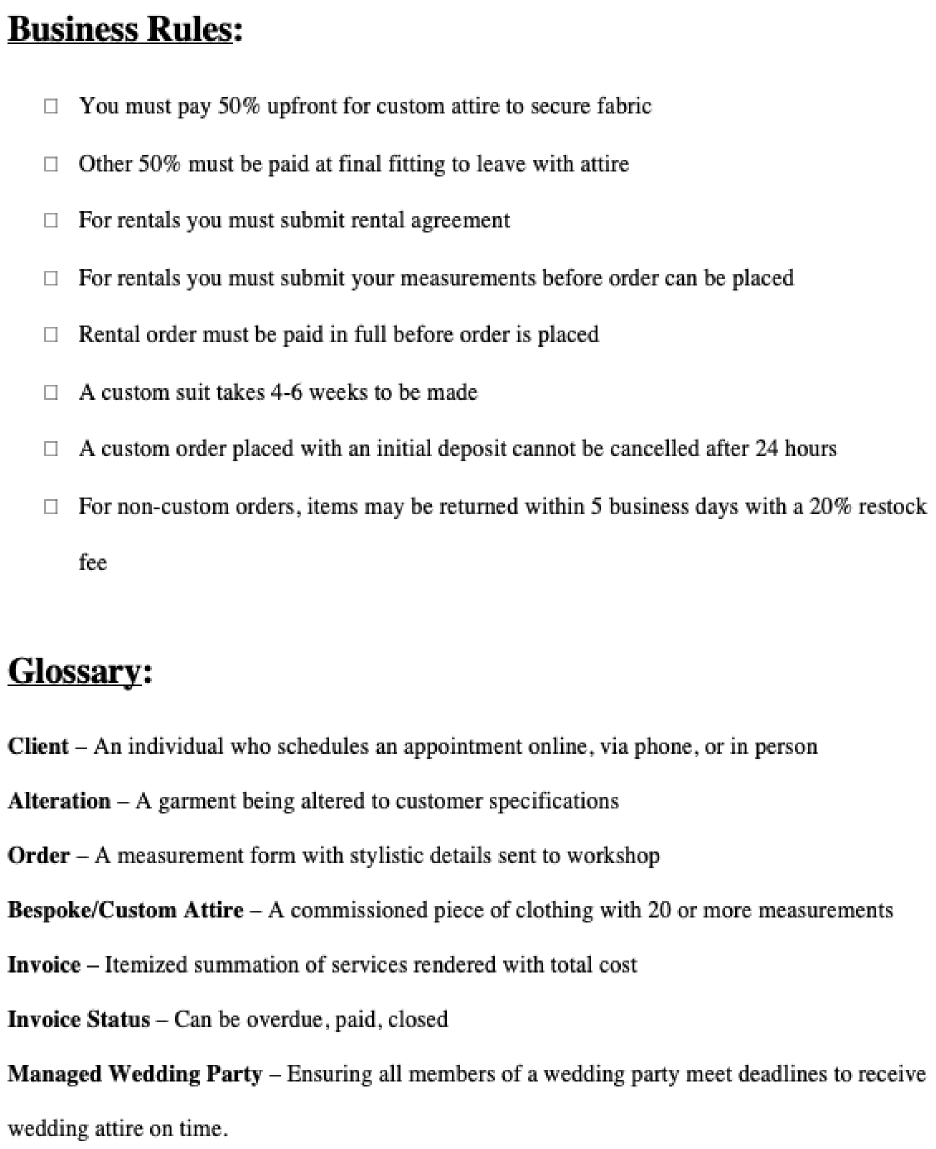
Project Description:

The Database Administration and Database Management course was facilitated by Professor Chad Harper. The deliverable for this course consisted of two parts ultimately meant to culminate in the design and implementation of a database with the ability to solve a data management issue.

I elected to address a commercial enterprise problem for a small business by the name of Capital Custom Clothiers in Annapolis, Maryland. Capital Custom Clothiers is a family-owned bespoke tailor shop based in Annapolis, Maryland. Since 2010 they have steadily carved out a niche within the Annapolitan bespoke market. As a family business of 3 members, responsibilities are starkly defined. However, there is a clear gap in the ability to analyze sales concisely through one centralized modality. Currently, Capital uses QuickBooks to keep account of payments made and received but this is primarily used as a compliance mechanism for tax purposes (Akonnor, "IST 659," 2022). The description of the objective can be best summarized as:

"The database we would like to implement would not only provide specific data about the most popular products, but also give insight as to which business segments are the most lucrative as well. We expect that with the increased ability to discern customer proclivities, we will be able to narrow business focus to the highest yielding outputs."

To adequately assist the intended stakeholders, it was pertinent to gather information on their business rules and create a standardized glossary for terminology. Once the information was organized, I was able to begin conceptualizing the relationships between tables within the database.



Business Rules:

□ You must pay 50% upfront for custom attire to secure fabric

□ Other 50% must be paid at final fitting to leave with attire

□ For rentals you must submit rental agreement

□ For rentals you must submit your measurements before order can be placed

□ Rental order must be paid in full before order is placed

□ A custom suit takes 4-6 weeks to be made

□ A custom order placed with an initial deposit cannot be cancelled after 24 hours

□ For non-custom orders, items may be returned within 5 business days with a 20% restock fee

Glossary:

Client – An individual who schedules an appointment online, via phone, or in person

Alteration – A garment being altered to customer specifications

Order – A measurement form with stylistic details sent to workshop

Bespoke/Custom Attire – A commissioned piece of clothing with 20 or more measurements

Invoice – Itemized summation of services rendered with total cost

Invoice Status – Can be overdue, paid, closed

Managed Wedding Party – Ensuring all members of a wedding party meet deadlines to receive wedding attire on time.

Figure 1: Business Rules & Glossary, (Akonnor, "IST 659," 2022)

The next phase of the database creation was the end-to-end creation of a conceptual model to a fully normalized logical model in third normal form. In the final stage of the logical model, the redundant relationships were decomposed and removed leaving only normalized relation between the primary keys and foreign keys of each table. Attributes were concretized to optimize the businesses' ability to analyze data for enhanced customer insights.

ERD - Conceptual Model:

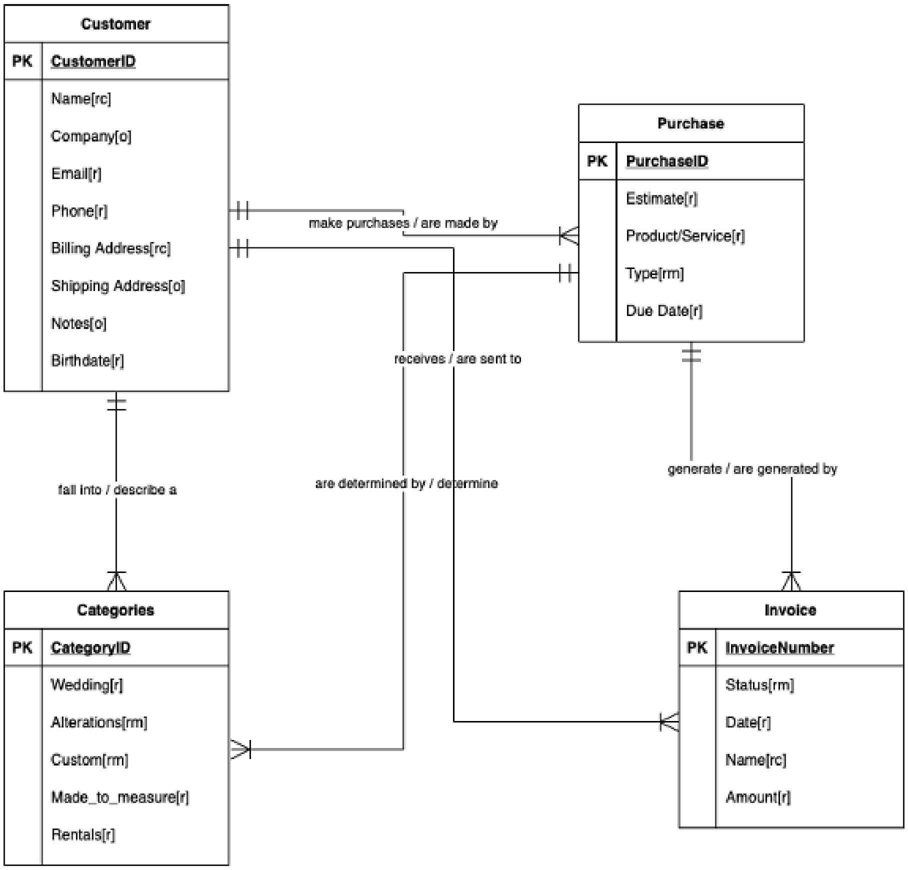


Figure 2: Entity-Relationship Diagram, (Akonnor, "IST 659," 2022)

Logical Model, Deliverable 1:

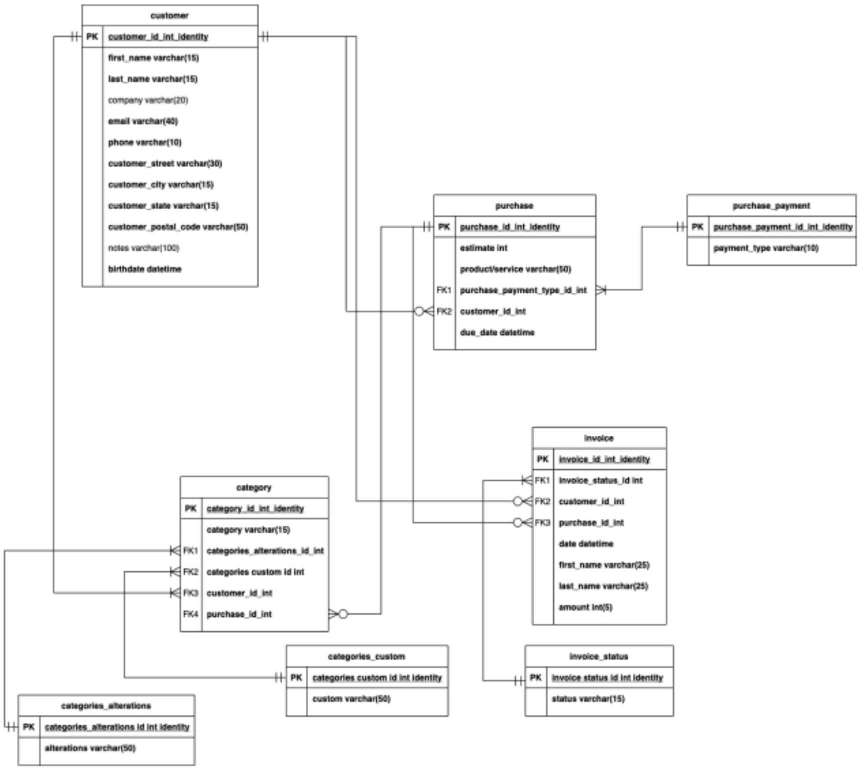


Figure 3: Logical Model 1, (Akonnor, "IST 659," 2022)

Logical Model, Deliverable 2:

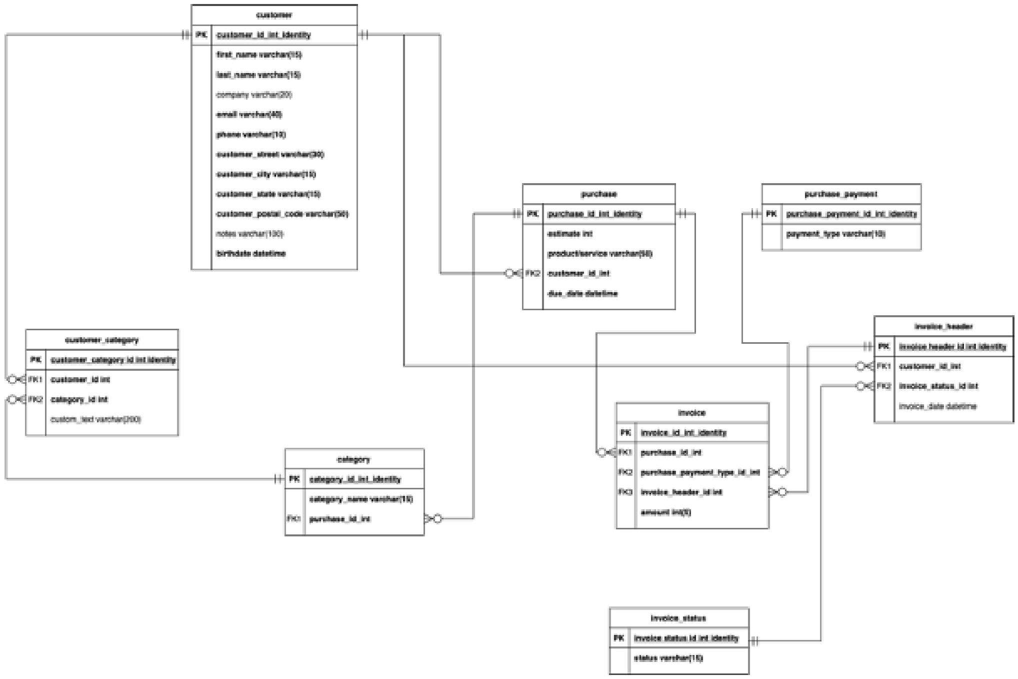
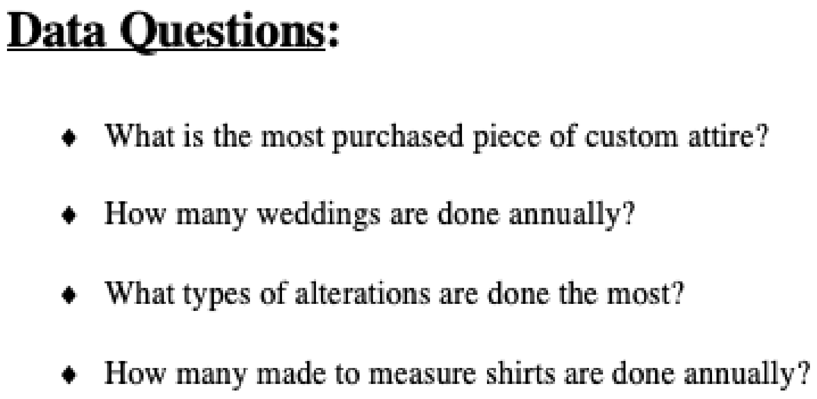

customer
PK customer_id_int_identity
first_name varchar(15)
last_name varchar(15)
company varchar(20)
email varchar(40)
phone varchar(10)
customer_street varchar(30)
customer_city varchar(15)
customer_state varchar(15)
customer_postal_code varchar(50)
notes varchar(100)
birthdate datetime
purchase
PK purchase_id_int_identity
estimate int
product/service varchar(50)
FK2 customer_id_int
due_date datetime
purchase_payment
PK purchase_payment_id_int_identity
payment_type varchar(10)
invoice_header
PK invoice header id int identity
FK1 customer_id_int
FK2 invoice_status_id int
invoice_date datetime
customer_category
PK customer_category id int identity
FK1 customer_id int
FK2 category_id int
custom_text varchar(200)
invoice
PK invoice_id_int_identity
FK1 purchase_id_int
FK2 purchase_payment_type_id_int
FK3 invoice_header_id int
amount int(5)
category
PK category_id_int_identity
category_name varchar(15)
FK1 purchase_id_int
invoice_status
PK invoice status id int identity
status varchar(15)

Figure 4: Logical Model 2, (Akonnor, "IST 659," 2022)

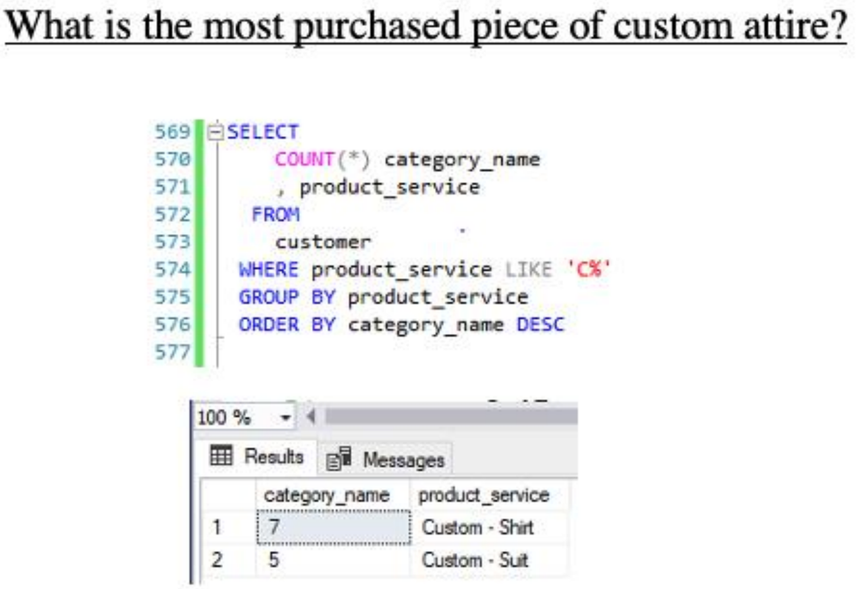
The last phase of this project consisted of converting the final logical model into a physical database using SQL in SQL Server Management studio. As a test of efficacy, the database needed to be capable of answering specific data business questions. An acknowledgement, for ethical reasons, mock data was used to protect the personally identifiable information of Capital Customer Clothier clients such as home address, credit card information, phone number, etc.



Data Questions:

- What is the most purchased piece of custom attire?
- How many weddings are done annually?
- What types of alterations are done the most?
- How many made to measure shirts are done annually?

Figure 5: Data Questions, (Akonnor, "IST 659," 2022)



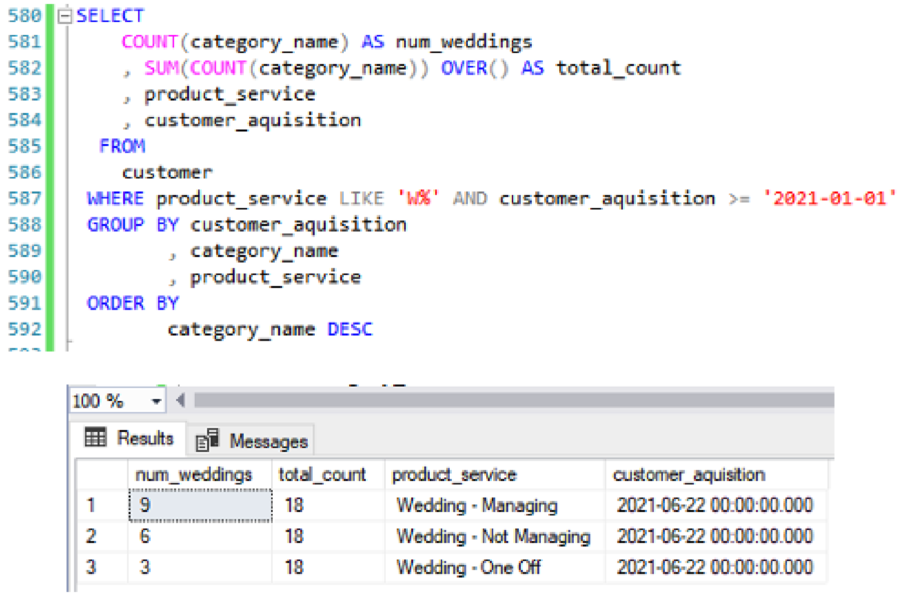
What is the most purchased piece of custom attire?

```
SELECT
    COUNT(*) category_name
    , product_service
  FROM
    customer
 WHERE product_service LIKE 'C%'
 GROUP BY product_service
 ORDER BY category_name DESC
```

	category_name	product_service
1	7	Custom - Shirt
2	5	Custom - Suit

Figure 6: Data Question 1, (Akonnor, "IST 659," 2022)

How many weddings are done annually?

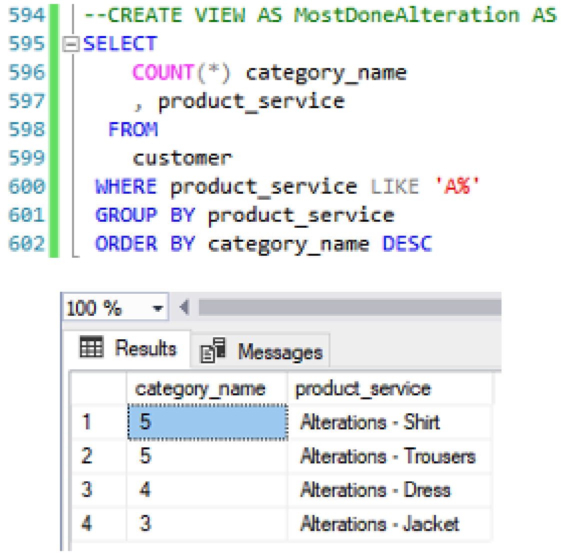


```
SELECT
    COUNT(category_name) AS num_weddings
    , SUM(COUNT(category_name)) OVER() AS total_count
    , product_service
    , customer_aquisition
  FROM
    customer
 WHERE product_service LIKE 'W%' AND customer_aquisition >= '2021-01-01'
 GROUP BY customer_aquisition
        , category_name
        , product_service
 ORDER BY
        category_name DESC
```

	num_weddings	total_count	product_service	customer_aquisition
1	9	18	Wedding - Managing	2021-06-22 00:00:00.000
2	6	18	Wedding - Not Managing	2021-06-22 00:00:00.000
3	3	18	Wedding - One Off	2021-06-22 00:00:00.000

Figure 7: Data Question 2, (Akonnor, "IST 659," 2022)

What types of alterations are done the most?

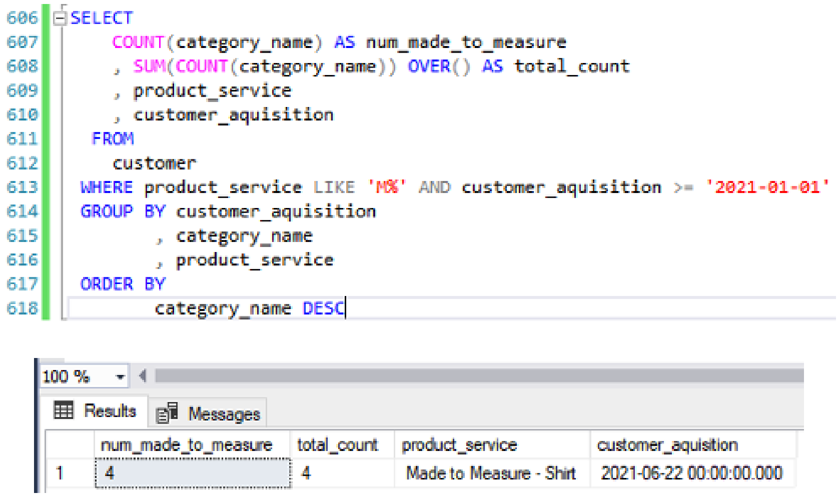


```
--CREATE VIEW AS MostDoneAlteration AS
SELECT
    COUNT(*) category_name
    , product_service
  FROM
    customer
 WHERE product_service LIKE 'A%'
 GROUP BY product_service
 ORDER BY category_name DESC
```

	category_name	product_service
1	5	Alterations - Shirt
2	5	Alterations - Trousers
3	4	Alterations - Dress
4	3	Alterations - Jacket

Figure 8: Data Question 3, (Akonnor, "IST 659," 2022)

How many made to measure shirts are done annually?



```
SELECT
    COUNT(category_name) AS num_made_to_measure
    , SUM(COUNT(category_name)) OVER() AS total_count
    , product_service
    , customer_aquisition
  FROM
    customer
 WHERE product_service LIKE 'M%' AND customer_aquisition >= '2021-01-01'
 GROUP BY customer_aquisition
        , category_name
        , product_service
 ORDER BY
        category_name DESC
```

	num_made_to_measure	total_count	product_service	customer_aquisition
1	4	4	Made to Measure - Shirt	2021-06-22 00:00:00.000

Figure 9: Data Question 4, (Akonnor, "IST 659," 2022)

The above are example of the querying power of the created database for the business. Should the stakeholders want the necessity for SQL experience abstracted away, Microsoft Access can be implemented on top of it. Overall, the small addition of a database created previously untapped analytic capabilities as well as now being able to inform business operational decisions.

In the nascent stages of my Data Science journey this project certainly felt like a herculean task. As time progressed and I slowly gleaned more context and knowhow regarding databases and the requisite normalization. I was able to collect and organize data regarding operational structure which later informed my plans for creating a functional database for the business. The answers to the business questions informed future decision making by illuminating purchase patterns that were previously subject to instinct alone.

Project 2:

IST 652: Scripting for Data Analysis
Professor: Deborah V. Landowski, PhD

Project Description:

The Scripting for Data Analysis course was facilitated by Deborah V. Landowski, PhD. The learning objectives were meant to grant students the ability to acquire, access, and transform data whether it existed in structured, semi-structured, or unstructured forms. The deliverable, "LA & NYC, A Decade of Crime Comparison", executed on the outlined objectives by conducting a crime analysis of New York and Los Angeles from disparate data sets and included a sentiment analysis from Twitter using API calls:

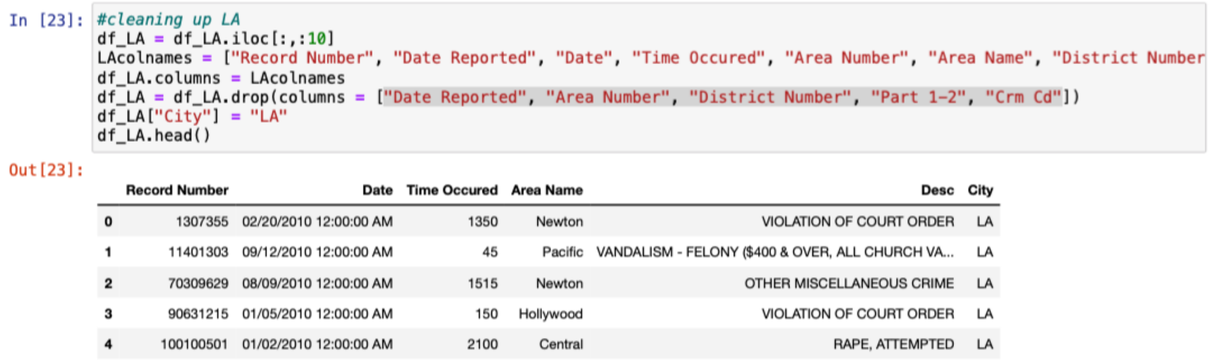
"The selected data sets from Kaggle, LA Crime Data and NYC Crime Data, provide longitudinal examinations of crime trends over a decade. A greater perspective can be gleaned regarding the types of crimes committed, the amount of crime committed, and what time periods crime was most prevalent between two major cities on opposite coasts. The intention, a comparative analysis demonstrating either a parallel or inverse relationship with crime patters. These two cities are amongst the largest and most densely populated urban areas in America which make them more prone to visible shifts in crime and the ideal candidates for examination (Akonnor, "IST 652," 2022)."

Once the data was collected, we sought out to answer 5 key questions:

Question 1: When was crime most prevalent in both cities?
Question 2: What were the top 10 most common crimes across both cities?
Question 3: When did crime in LA peak?
Question 4: What was the most common crime in LA in 2014?
Question 5: Crime, specifically burglary, was at its highest in 2017. What was the conversation on Twitter like during 2017 in LA regarding burglary? (Akonnor, "IST 652," 2022)."

With these research findings leading our inquiry, the analysis with Jupyter Notebook in Python3 commenced. The order of operations were the datasets being cleaned and merged, comparison questions presented, data mining, patterns identified and visualized, and an analysis of the findings.

The data cleaning was executed by importing the individual CSV files into the notebook. The first data set, LA Crime Data, had 2,118,203 rows and 28 columns. After the initial import, the data was soon reduced to the first 10 columns with the row count remaining intact. The final data frame for analysis consisted of 2,118,203 rows and 5 columns excluding the index (Akonnor, “IST 652,” 2022).



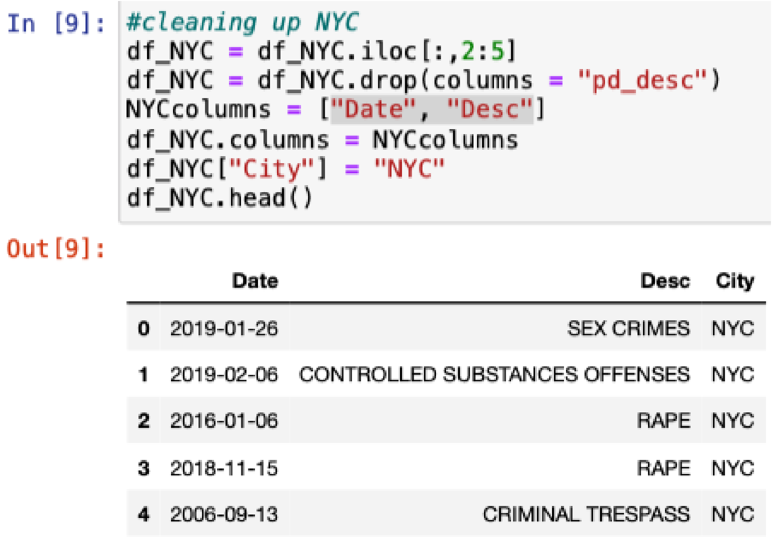
```
In [23]: #cleaning up LA
         df_LA = df_LA.iloc[:,:10]
         LAcolnames = ["Record Number", "Date Reported", "Date", "Time Occured", "Area Number", "Area Name", "District Number
         df_LA.columns = LAcolnames
         df_LA = df_LA.drop(columns = ["Date Reported", "Area Number", "District Number", "Part 1-2", "Crm Cd"])
         df_LA["City"] = "LA"
         df_LA.head()
```

Out[23]:

	Record Number	Date	Time Occured	Area Name	Desc	City
0	1307355	02/20/2010 12:00:00 AM	1350	Newton	VIOLATION OF COURT ORDER	LA
1	11401303	09/12/2010 12:00:00 AM	45	Pacific	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	LA
2	70309629	08/09/2010 12:00:00 AM	1515	Newton	OTHER MISCELLANEOUS CRIME	LA
3	90631215	01/05/2010 12:00:00 AM	150	Hollywood	VIOLATION OF COURT ORDER	LA
4	100100501	01/02/2010 12:00:00 AM	2100	Central	RAPE, ATTEMPTED	LA

Figure 10: LA Dataset Cleaning, (Akonnor, “IST 652,” 2022)

The second data set, NYC Crime Stats, underwent a similar revision. Initially, the data set began with 3,881,989 rows and 19 columns. After the import, the data was soon reduced to columns 2-5 with the row count remaining intact (Akonnor, “IST 652,” 2022).



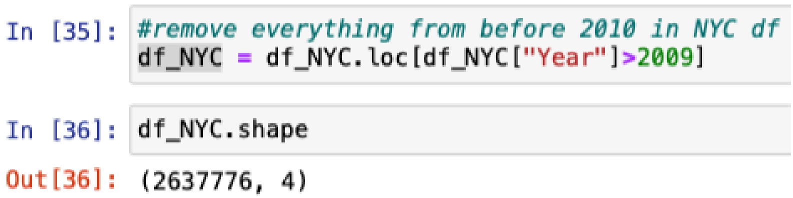
```
In [9]: #cleaning up NYC
        df_NYC = df_NYC.iloc[:,2:5]
        df_NYC = df_NYC.drop(columns = "pd_desc")
        NYCcolumns = ["Date", "Desc"]
        df_NYC.columns = NYCcolumns
        df_NYC["City"] = "NYC"
        df_NYC.head()
```

Out[9]:

	Date	Desc	City
0	2019-01-26	SEX CRIMES	NYC
1	2019-02-06	CONTROLLED SUBSTANCES OFFENSES	NYC
2	2016-01-06	RAPE	NYC
3	2018-11-15	RAPE	NYC
4	2006-09-13	CRIMINAL TRESPASS	NYC

Figure 11: NYC Dataset Cleaning, (Akonnor, “IST 652,” 2022)

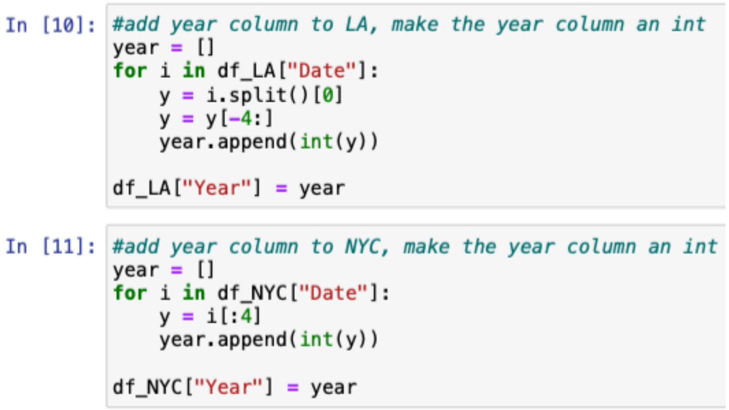
Since a decade is the time horizon being examined, all data prior to the year 2009 was removed. The final data frame for analysis consisted of 2,637,776 rows and 3 columns excluding the index.



```
In [35]: #remove everything from before 2010 in NYC df
         df_NYC = df_NYC.loc[df_NYC["Year"]>2009]

In [36]: df_NYC.shape

Out[36]: (2637776, 4)
```

Figure 12: NYC Dataset Cleaning 2, (Akonnor, "IST 652," 2022)

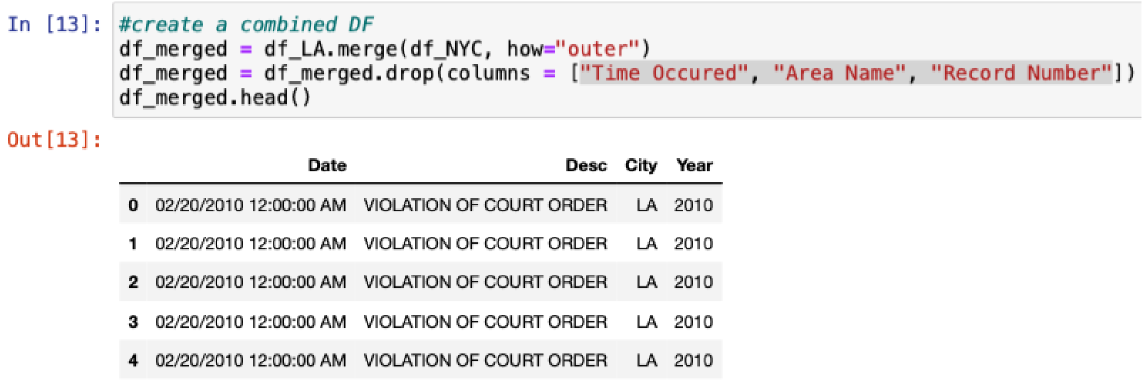
Once the data frames were cleaned and augmented to a satisfactory degree, an additional column, "Year", was added to the LA Crime Data and NYC Crime Stats respectively.



```
In [10]: #add year column to LA, make the year column an int
         year = []
         for i in df_LA["Date"]:
             y = i.split()[0]
             y = y[-4:]
             year.append(int(y))

         df_LA["Year"] = year

In [11]: #add year column to NYC, make the year column an int
         year = []
         for i in df_NYC["Date"]:
             y = i[:4]
             year.append(int(y))

         df_NYC["Year"] = year
```

Figure 13: Column Addition, (Akonnor, "IST 652," 2022)

Finally, the data sets were combined into one merged data framed called "df_merged" by an outer join.



```
In [13]: #create a combined DF
         df_merged = df_LA.merge(df_NYC, how="outer")
         df_merged = df_merged.drop(columns = ["Time Occured", "Area Name", "Record Number"])
         df_merged.head()

Out[13]:
                     Date                      Desc  City  Year
   0  02/20/2010 12:00:00 AM  VIOLATION OF COURT ORDER    LA  2010
   1  02/20/2010 12:00:00 AM  VIOLATION OF COURT ORDER    LA  2010
   2  02/20/2010 12:00:00 AM  VIOLATION OF COURT ORDER    LA  2010
   3  02/20/2010 12:00:00 AM  VIOLATION OF COURT ORDER    LA  2010
   4  02/20/2010 12:00:00 AM  VIOLATION OF COURT ORDER    LA  2010
```

Figure 14: Merged Datasets, (Akonnor, "IST 652," 2022)

A functional dataset successfully created fertile ground for analysis. To accomplish this, the first program begins by importing the appropriate packages for analysis. In this instance the "csv" module is imported as well as "pandas" as "pd". This program answered the first 4 research questions

Question 1: When was crime most prevalent in both cities?

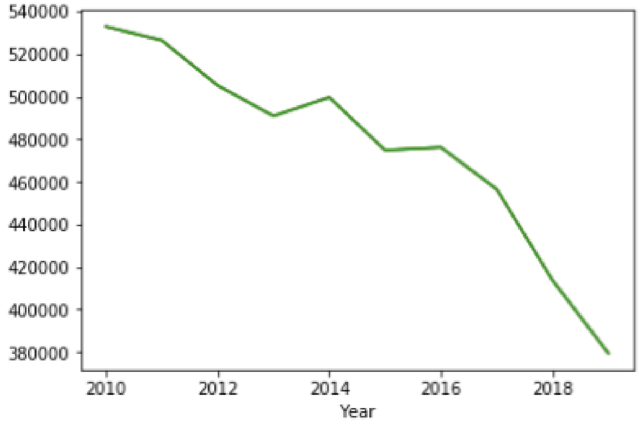
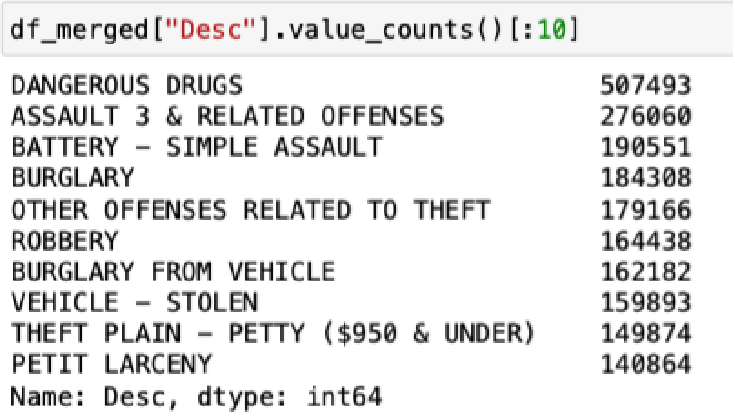


Figure 15: Crime Prevalence, (Akonnor, "IST 652," 2022)

The above plot showed a stark contrast in crime between 2010 and 2020 for the cities combined. In 2010, there were approximately 538,000 crimes ultimately proving to be the peak of crime prevalence in both cities. There was a declining trend with a brief disruption in 2014 where crime peaked, dropped shortly after, plateaued, and then followed a steady decline until 2020. Presumably the start of the pandemic.

Question 2: What were the top 10 most common crimes across both cities?



```
df_merged["Desc"].value_counts()[:10]
```

```
DANGEROUS DRUGS                          507493
ASSAULT 3 & RELATED OFFENSES             276060
BATTERY - SIMPLE ASSAULT                 190551
BURGLARY                                 184308
OTHER OFFENSES RELATED TO THEFT          179166
ROBBERY                                  164438
BURGLARY FROM VEHICLE                    162182
VEHICLE - STOLEN                         159893
THEFT PLAIN - PETTY ($950 & UNDER)       149874
PETIT LARCENY                            140864
Name: Desc, dtype: int64
```

Figure 16: Most Common Crimes, (Akonnor, "IST 652," 2022)

The largest count of crimes committed were classified as "Dangerous Drugs" at 507,493. This makes sense intuitively as the US has undergone various magnitudes of opioid epidemics. Drug related crimes tend to be more personal, isolation wise and can be committed in the privacy of a dwelling making them more likely to occur. The next two categories of crimes occupying spots two and three were "Assault 3 & Related Offenses" at 276,060 instances and "Battery – Simple Assault" at 190,551 instances respectively. The next two were expected in larger cities. People in proximity interact more frequently leading to more altercations. Interestingly, the next 7 crimes were all related to burglary/theft in some capacity. Germane to the most popular type of crime, theft is most often executed in isolation.

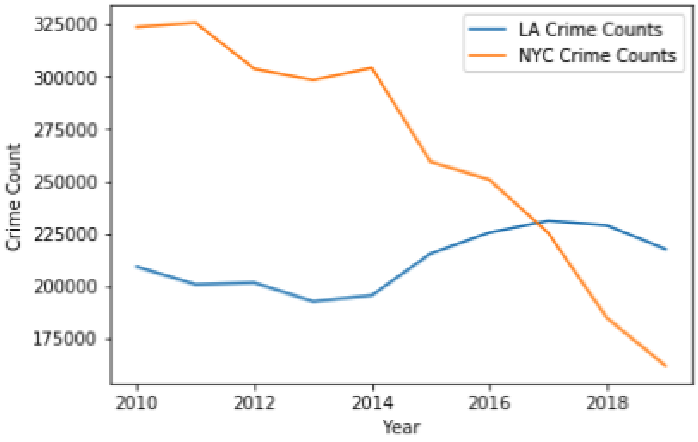


Figure 17: Separated Crime Trends, (Akonnor, "IST 652," 2022)

When the data was separated, it revealed that the initial line plot was heavily influenced by NYC's crime trends. LA in that decade had considerably less crime until around 2017 when the overall crime volume in LA exceeded that of NYC handedly. This unforeseen movement in the line was the impetus for a more in-depth look at when specially in the arch in occurred.

Question 3: When did crime in LA peak?

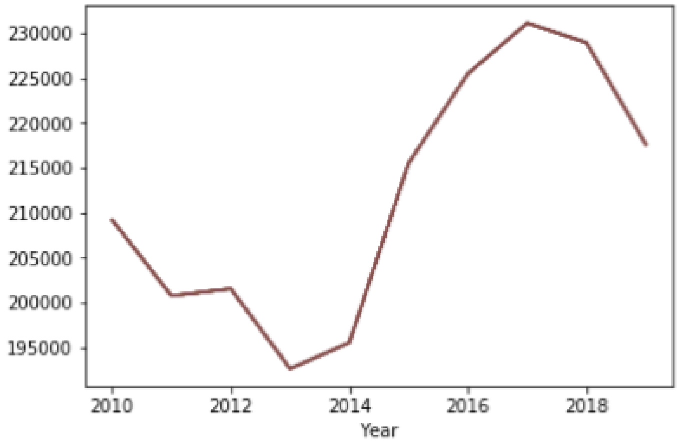
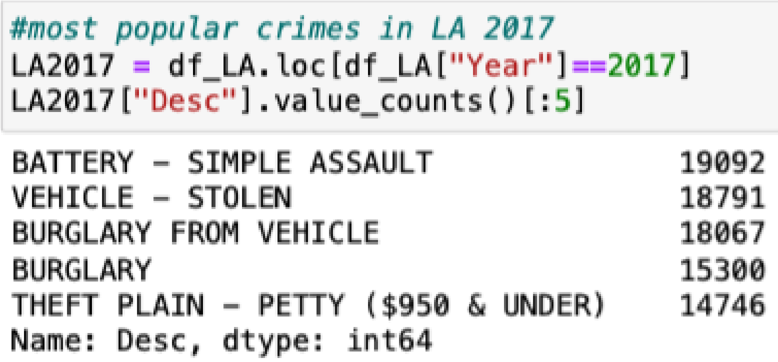


Figure 18: LA Crime Peak, (Akonnor, "IST 652," 2022)

The above plot allowed for a clearer examination of crime trends. It is apparent that crime jumped drastically from 2013-2017. In 2013 there were less than 195,000 crimes and that number rose alarmingly to 230,000 in 2017. Various sources at the time suggest that this was due to emboldened racial supremacist groups committing hate crimes toward Asians, who comprise a large portion of LA's population. All of this on the heels of Donald Trump's election to president at the time.

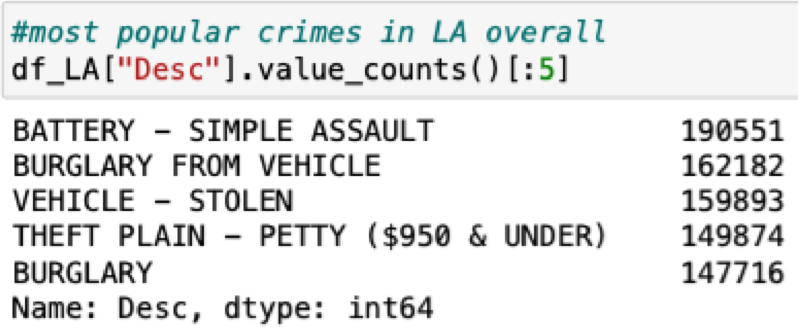


```
#most popular crimes in LA 2017
LA2017 = df_LA.loc[df_LA["Year"]==2017]
LA2017["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                19092
VEHICLE - STOLEN                        18791
BURGLARY FROM VEHICLE                   18067
BURGLARY                                15300
THEFT PLAIN - PETTY ($950 & UNDER)      14746
Name: Desc, dtype: int64
```

Figure 19: LA Most Popular Crimes, (Akonnor, "IST 652," 2022)

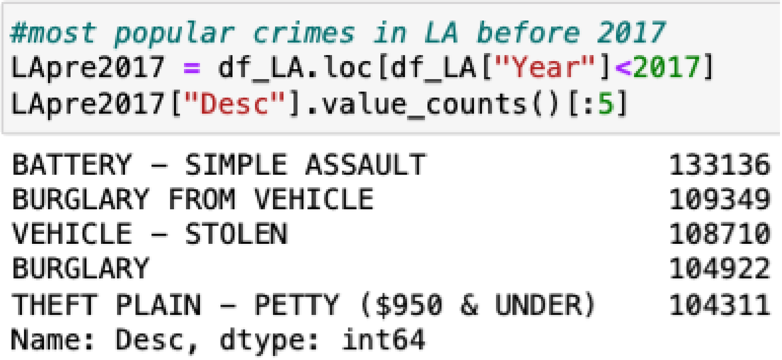
While this presupposition is a theory, the data suggests that there could be a correlation. When the LA data was broken down into the top 5 most highly occurring crimes in 2017, the number 1 crime was "Battery – Simple Assault." This alone without the context of racial motivation detailed in the crimes is not damning evidence but spurred further investigation.



```
#most popular crimes in LA overall
df_LA["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                 190551
BURGLARY FROM VEHICLE                    162182
VEHICLE - STOLEN                         159893
THEFT PLAIN - PETTY ($950 & UNDER)       149874
BURGLARY                                 147716
Name: Desc, dtype: int64
```

Figure 20: LA Most Popular Crime Overall, (Akonnor, "IST 652," 2022)



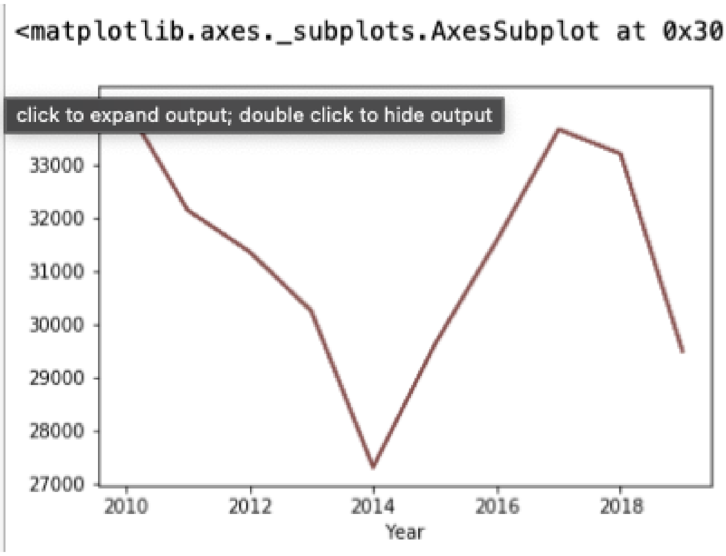
```
#most popular crimes in LA before 2017
LApre2017 = df_LA.loc[df_LA["Year"]<2017]
LApre2017["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                 133136
BURGLARY FROM VEHICLE                    109349
VEHICLE - STOLEN                         108710
BURGLARY                                 104922
THEFT PLAIN - PETTY ($950 & UNDER)       104311
Name: Desc, dtype: int64
```

Figure 21: Most popular crimes in LA prior to 2017, (Akonnor, "IST 652," 2022)

If the spike were to be attributed to a shift in political power, then it could stand to reason assault would have moved into the number 1 spot from a lower position. When the totality of crimes in LA was observed, it showed that "Battery – Simple Assault" was still in the first spot by 30,000 instances with a total of 190,551. This number suggests that assaults have simply historically been the most committed crime in LA, but a fair counter argument could have been that the number was the result of the massive 2017 outlier. The next step was to appraise assaults prior to 2017. The data still revealed a 24,000-instance gap between assault and the next most common crime of "Burglary from Vehicle." In summation, though it is possible to attribute the 2017 spike to political instigation, more and specific data would be needed to make any concrete conclusions.

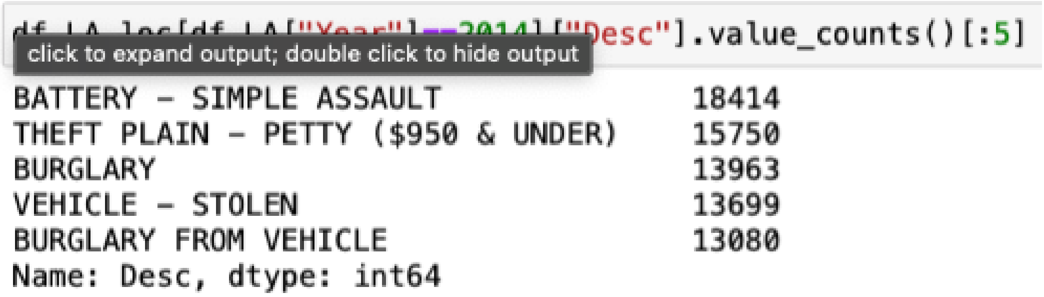
Question 4: What was the most common crime in LA in 2014?



```
<matplotlib.axes._subplots.AxesSubplot at 0x3016ac9b0>
```

Figure 22: Most popular crimes in LA prior to 2017, (Akonnor, "IST 652," 2022)

Moving in the opposite direction, 2014 was one of the lowest crime points in the decade for crime in LA. This piqued an inquiry into if the low may have also influenced a shift in the ordinal placement of the most common crimes over the decade.



```
df_LA.loc[df_LA["Year"]==2014]["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                18414
THEFT PLAIN - PETTY ($950 & UNDER)      15750
BURGLARY                                13963
VEHICLE - STOLEN                        13699
BURGLARY FROM VEHICLE                   13080
Name: Desc, dtype: int64
```

Figure 23: Most popular crimes in LA Descending (Akonnor, "IST 652," 2022)

Even in 2014, the order and crimes remained consistent with "Battery – Simple Assault" still at the top. This solidified that it was the most committed crime in LA during the last decade. However, if all iterations of theft/burglary were combined, they comprised a massive portion of the pie.

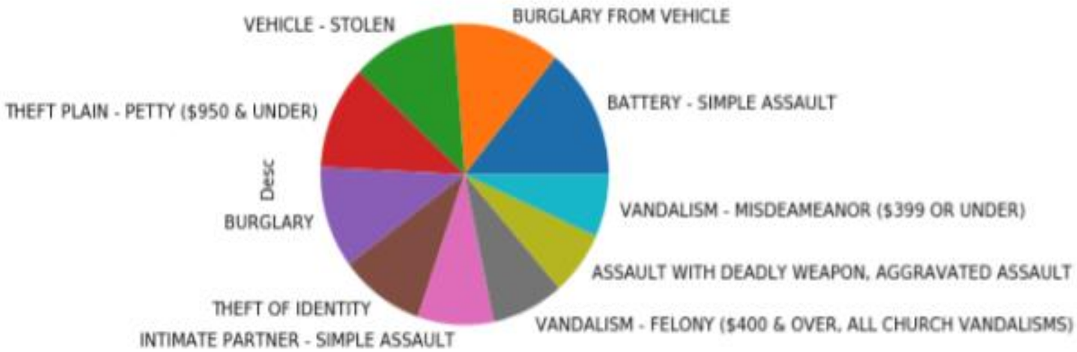
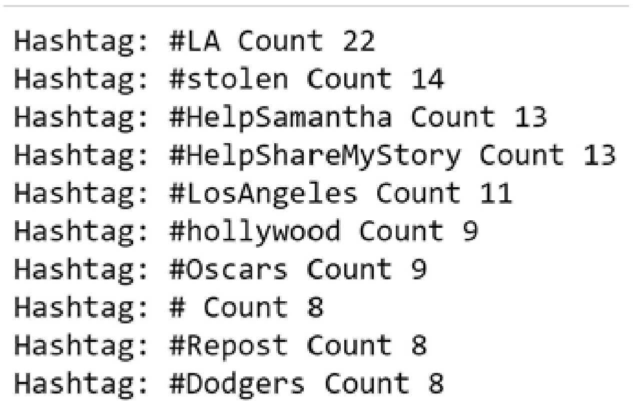


Figure 24: Crime Pie Chart, (Akonnor, "IST 652," 2022)

Question 5: Crime, specifically burglary, was at its highest in 2017. What was the conversation on Twitter like during 2017 in LA regarding burglary?



```
Hashtag: #LA Count 22
Hashtag: #stolen Count 14
Hashtag: #HelpSamantha Count 13
Hashtag: #HelpShareMyStory Count 13
Hashtag: #LosAngeles Count 11
Hashtag: #hollywood Count 9
Hashtag: #Oscars Count 9
Hashtag: # Count 8
Hashtag: #Repost Count 8
Hashtag: #Dodgers Count 8
```

Figure 25: Crime Hashtags, (Akonnor, "IST 652," 2022)

Hashtags are often used thematically at the end of a tweet to give it a categorization. While useful with full context, without it, having a proper understanding of what's going on was quite difficult to ascertain. Above were the top 10 hashtags for tweets containing the words "burglary", "burglar", "burglarized", "stolen" and or "stole" in the year 2017. The insinuation is that perhaps 3 of the hashtags "stolen", "Help Samantha", and "HelpShareMyStory" pertain to burglary in the manner intended. The other 7, seemingly are the opinions of sports or movie fans who believe their team, or celebrity was robbed of a rightful victory.

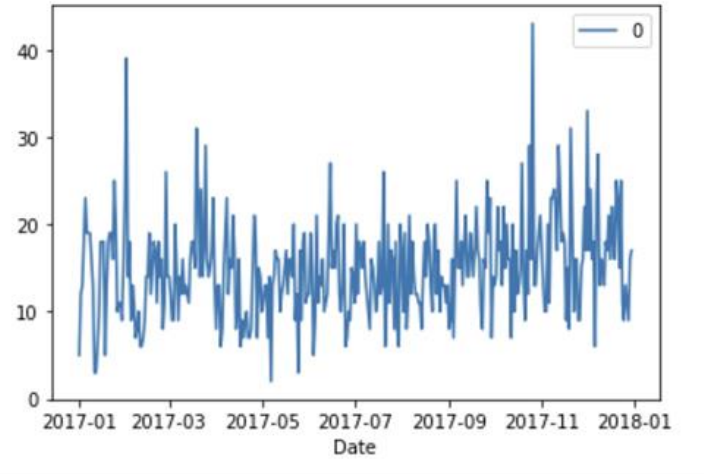
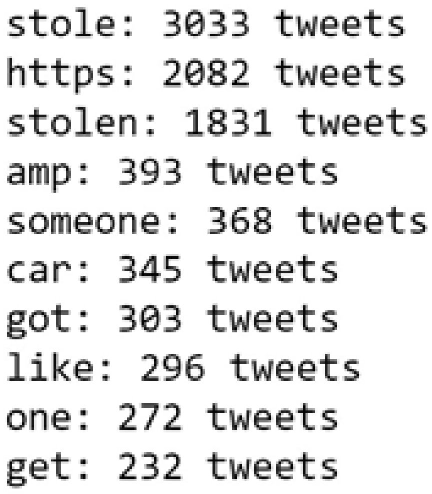


Figure 26: Time series, (Akonnor, "IST 652," 2022)

The above line plot shows spikes in conversations pertaining to theft over the one-year time horizon from 2017 to 2018. January and November had the most noticeable spikes up to 40 with the median density around 18 conversations at any given time.



```
stole: 3033 tweets
https: 2082 tweets
stolen: 1831 tweets
amp: 393 tweets
someone: 368 tweets
car: 345 tweets
got: 303 tweets
like: 296 tweets
one: 272 tweets
get: 232 tweets
```

Figure 27: Top 10 Keywords, (Akonnor, "IST 652," 2022)

Finally, the top 10 keywords. Unsurprisingly the word "stole" is first with 3033 tweets, "stolen" is in third place with 1831, and the rest save the word "car" began to get somewhat arbitrary.

Initial curiosity about crime in densely populated urban areas was the impetus of this project. The further pursuit of this topic led to the acquisition of two disparate datasets intended for the use of comparative analysis. To properly conduct the analysis, the data required cleaning and concatenation through manipulation in Python. Once a fully functional dataset was constructed, exploration of the data revealed patterns and trend spurring 5 research questions. These questions were answered with the assistance and various data plotting methods. Per the stated goals of IST 652 and the program, the learning objectives were accomplished through the execution of this project.

Project 3:

IST 718: Big Data Analytics

Professor: Jon Fox, PhD

Project Description:

The Big Data Analytics course was facilitated by Professor Jon Fox, PhD. The learning objectives of the project were meant to solidify the OSEMN framework that was continually reinforced throughout the course. The deliverable, a comprehensive business report that addressed the inquiry of internal stakeholders regarding appropriate insurance premiums over the next 5 years for the company. This report was comprised of average temperature forecasts and an examination of insurance premium trends over a 21-year span. Our team, the analytics division of InsurTex, executed on the outlined objectives by obtaining data, scrubbing it, exploring the datasets, modeling the data, and interpreting the results.

The project commenced with data collection from a plethora of sources such as Corporate Insurance Data from 2008-2021, FEMA National Risk Index, Insurance Information Institute, Energy Star – Texas Weather, and many more. Not all sources were used for analysis however, they served to provide context and direction for further research. The data was ingested into a Jupyter notebook for analysis with various Pandas libraries as well as scrubbed and concatenated.

Once complete, the data was prepared for exploration. City level datasets of average temperatures were created for the 7 cities of focus: El Paso, Amarillo, Dallas, Houston, Austin, Brownsville, and Laredo. These datasets were then overlaid for a visual comparison.

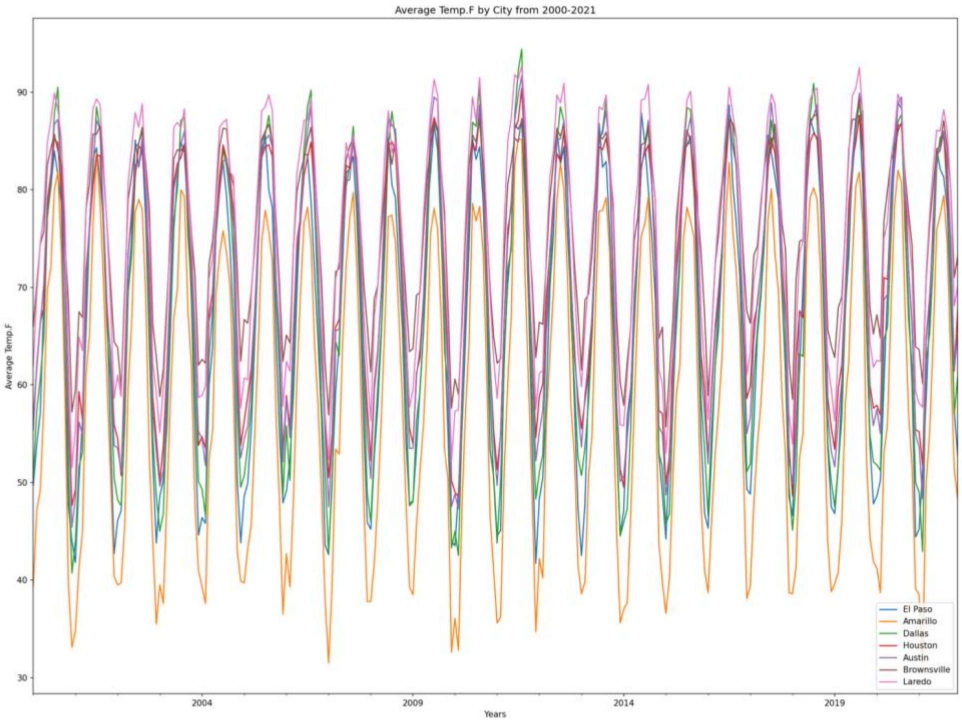


Figure 28: Average Temperature by City, (Akonnor, "IST 718," 2022)

The above plot displays average temperature by city from the years 2000 to 2021. Though Dallas had the highest average temperature over the 21-year time horizon with a spike to ~92 degrees in 2012, Loredo has consistently been the hottest of the 7 cities. Conversely, Amarillo has consistently had the lowest average temperatures without competition dropping to temperatures as frigid as ~30 degrees in 2007 and achieving it once again in 2021 (Akonnor, "IST 652," 2022). The data was then separated into average temperature charts by city for a more in-depth view. From there, projections were made for 48 months ahead to potentially mitigate the risk of upcoming extreme weather conditions.

This exploration revealed that there were no consistent average temperature patterns emerging globally amongst these 7 cities. Each was subject to its own fluctuations. This was illuminating from the perspective of not being able to ascribe a general outlook but, halting in that the need to proceed to the next data source became apparent. The insurance data revealed historical trends of insurance premiums rising year-over-year for the last 21 years and the forecasts corroborated these findings by also projecting increases for the next 5 fiscal years.

The final business recommendations were based on the final step of the analysis. The Texas Homeowners Insurance dataset was cleaned in preparation for use by the Prophet model. Below, a projection of steadily increasing insurance rates beginning in 2000 and continuing into the foreseeable future. This paralleled what the red trend line showed as well. There was a change point denoted by the perforated red line indicating a statistically significant shift occurring in 2014. The predicted values (black dots) were all very close to the actual data points (blue line) with limited variance.

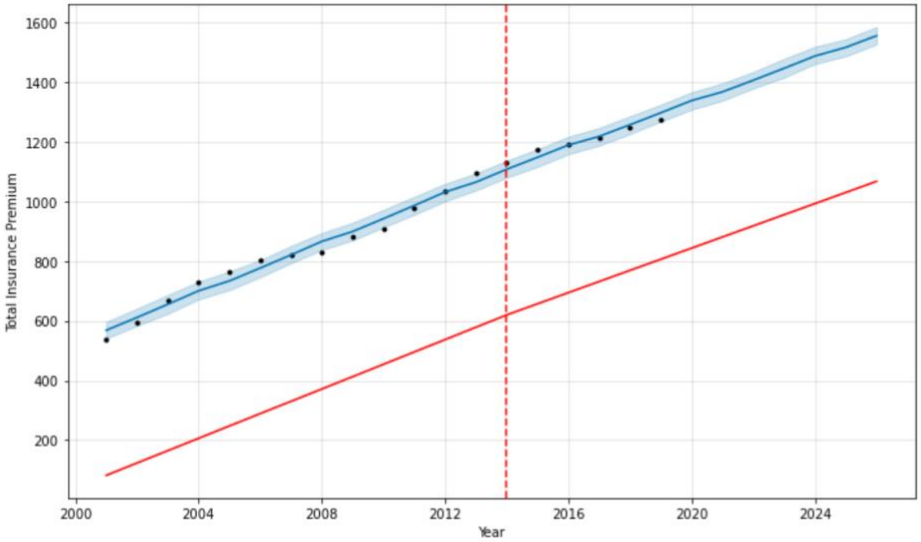


Figure 29: Insurance Premiums, (Akonnor, "IST 718," 2022)

The trendline was further isolated to emphasize the pattern of increasing rates over time. The third chart below showed the variance over a fiscal year. The fluctuations remained consistent for 11 months out of the year. Considering this was financial data, there was a drastic change in January when rates changed for the new year.

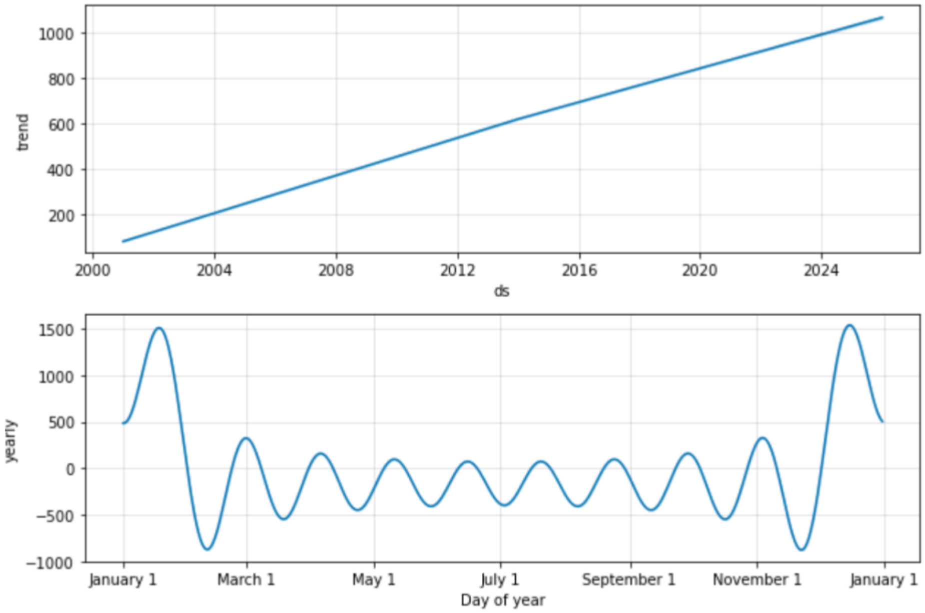


Figure 30: Trendlines, (Akonnor, "IST 718," 2022)

Our final premium recommendations for our company, InsurTex, for the next 5 years was as follows.

Fiscal Year	Premium Price
2022-2023	**$1,406.10**
2023-2024	**$1,446.09**
2024-2025	**$1,487.14**
2025-2026	**$1,516.30**
2026-2027	**$1,555.24**

Figure 31: Projected Premiums, (Akonnor, "IST 718," 2022)

Conclusion:

The contents of this portfolio are meant to be indicative of mastery regarding the four pillars of the Applied Data Science program: data collection, data analysis, strategy and decisions, and implementation. In the 3 presented projects, data was collected through various methodologies ranging from databases, csv files, and web scraping via APIs. The collection of such data was for the purpose of in-depth statistical analysis and data mining applications such as linear regression, logistic regression, and classification. These projects were executed with the express purpose of attaining an elite data science skillet through the guidance of exemplary faculty. These abilities provide ubiquitous problem-solving methodologies with application for any business problem, domain agnostic. Additionally, empowering me to communicate my findings to stakeholders with varying technical proficiency and provide sound and ethical courses of action.

I want to conclude by giving my immense gratitude to every educator, advisor, guest lecturer, and student I have had the privilege of interacting with over the last 18 months. Each one of you, no matter the scale, has contributed in tangible and imperceptible ways to create an individual who is data driven, ethically conscientious, passionate about problem solving, and excited to make an impact in the world. Thank you.

References:

Akonnor, M. (n.d.). *MSDS-Portfolio/CapitalCustomClothiers_Database_Project.pdf at main · Makonnor1/MSDS-Portfolio*. GitHub. Retrieved August 2, 2022, from https://github.com/Makonnor1/MSDS-Portfolio/blob/main/CapitalCustomClothiers_Database_Project.pdf

Akonnor, M. (n.d.-b). *MSDS-Portfolio/InsurTex - Final Project.pdf at main · Makonnor1/MSDS-Portfolio*. GitHub. Retrieved August 2, 2022, from https://github.com/Makonnor1/MSDS-Portfolio/blob/main/InsurTex%20-%20Final%20Project.pdf

Akonnor, M. (n.d.-c). *MSDS-Portfolio/LA & NYC - A DECADE OF CRIME COMPARISON.pdf at main · Makonnor1/MSDS-Portfolio*. GitHub. Retrieved August 2, 2022, from https://github.com/Makonnor1/MSDS-Portfolio/blob/main/LA%20%26%20NYC%20-%20A%20DECADE%20OF%20CRIME%20COMPARISON.pdf