Tasks

1. 1 sheet - 2 pages (technology, data set, findings)
2. Strong hypothesis, research, conclusion - viability, relevance, or usefulness of the technology.
3. References and Primary research.
4. Choose a Python or SQL tool, technique, or library that we didn't study in class.

Evaluation

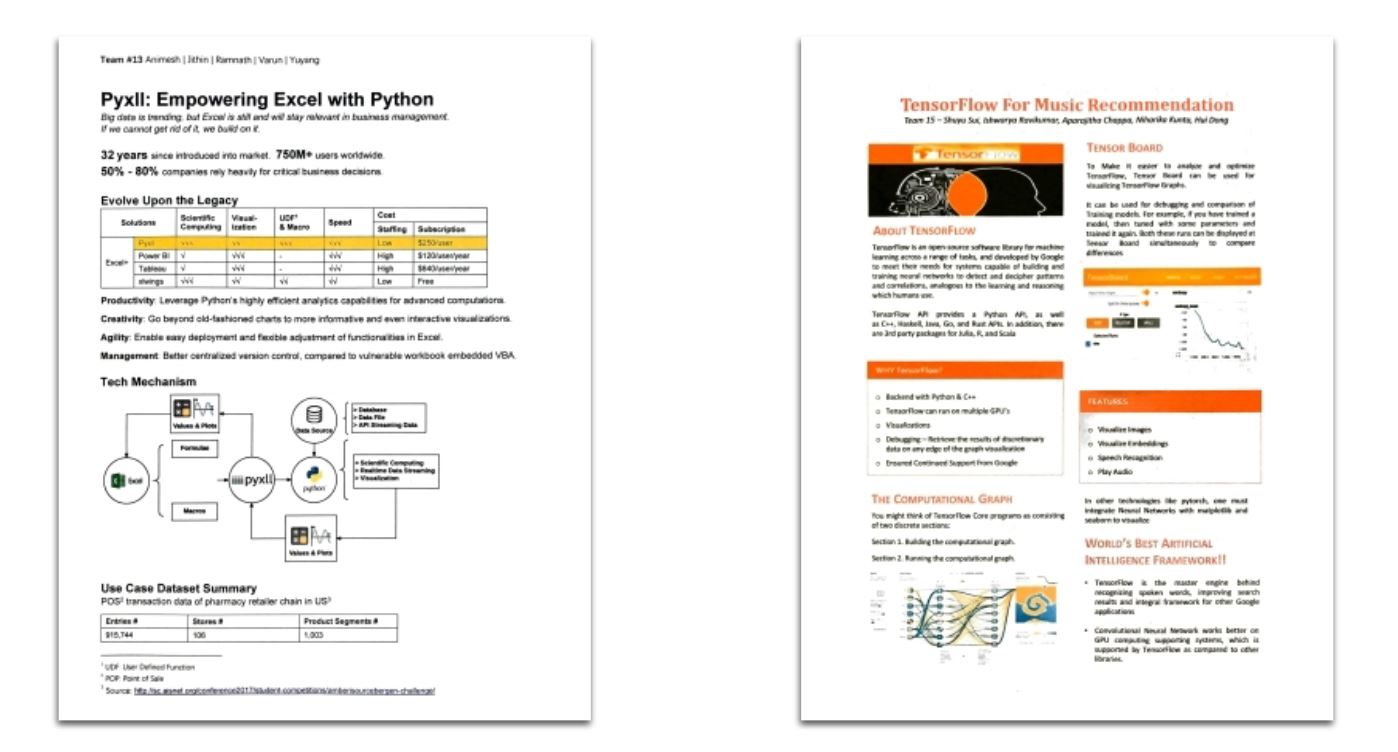
1. Novelty - Avoid similar topics.
2. Business Value - why an organization should adopt the technology
3. Briefing - short and to-the-point
4. Handout - eye-catching and informative

Our preference - Business value over Novelty.

Link - <https://www.gartner.com/document/3996980?ref=solrAll&refval=308220451>

Topic - "Don't have enough data? Small and wide data has your back!" (using Python)

Sample Flyer



Good article highlighting different techniques

<https://searchenterpriseai.techtarget.com/feature/Using-small-data-sets-for-machine-learning-models-sees-growth>

Datasets relevant to small and wide data

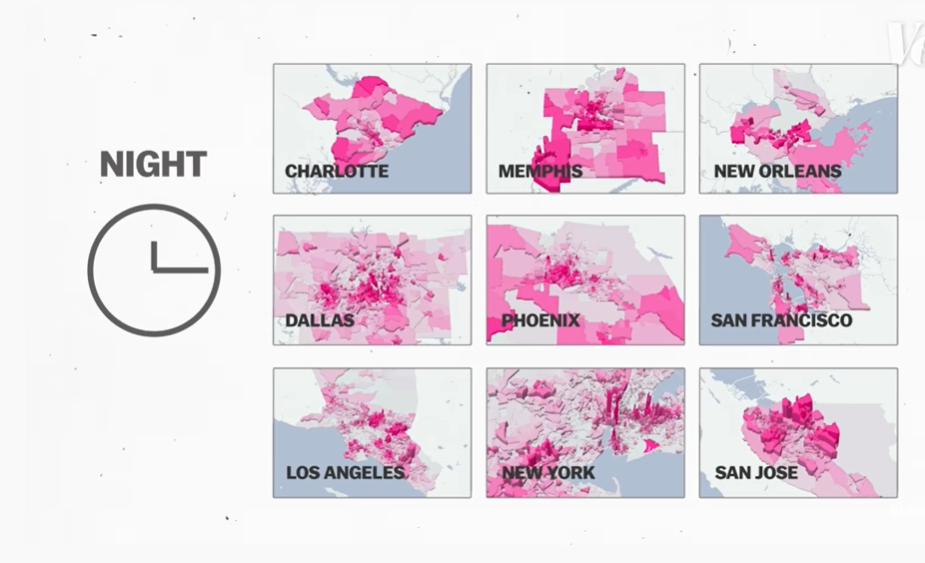
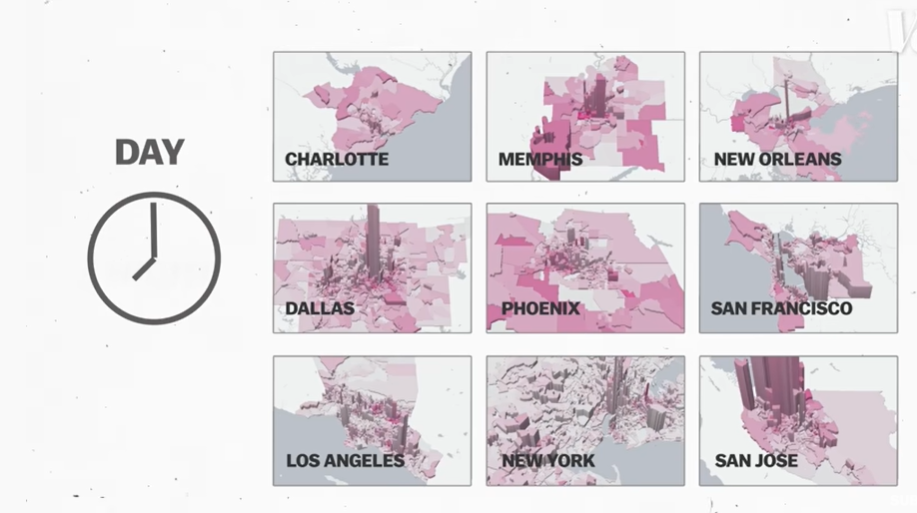
1. NYC Vehicle collisions and crashes (this would have a great visual appeal - I recollect seeing something similar in an NYT blog a while ago {NYT has amazing visualisations}) - <https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95>
2. NYC Payroll data - <https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e>

The below two datasets are quite comprehensive and have been presented as a web page instead of a single table (so we can cherry pick our data of interest), these are far more preferable for presentation but I'm unsure if we could consider it as small data (although we can reduce the data points for our purpose) it’s quite wide for sure.

1. NBA - <https://www.basketball-reference.com/>
2. MLB - <https://www.baseball-reference.com/>

Ideas for visuals in our flyer

1. <https://www.pinterest.com/skystage/nyt/>
2. The traffic collision data can be represented in some form of the below visuals.



References relevant for presentation/flyer -

1. <https://www.gartner.com/en/newsroom/press-releases/2021-05-19-gartner-says-70-percent-of-organizations-will-shift-their-focus-from-big-to-small-and-wide-data-by-2025>
2. <https://www.researchgate.net/publication/282158923_Small_data_in_the_era_of_big_data>
3. <https://www.gartner.com/document/3996980?ref=solrAll&refval=308920332>

Advantages of Small and wide data

1. Fast, easy, and cheaper

Typically - big data is when it’s unstructured and/or distributed across multiple data sources (typically having terabytes of data)

Idea - Suggest hypothesis for traffic data

1. Dataset - NY Traffic collisions dataset
2. Small data approach
   1. Visualization - usings some form of a map (primarily for visual appeal)
   2. Regression (Logistic/Multiple), Decision trees (might be interesting)
   3. Time series analysis
3. Wide data approach - <https://data.ny.gov/>
   1. Combine additional data points to build insights
      1. Real time traffic-speed data <https://data.cityofnewyork.us/Transportation/Real-Time-Traffic-Speed-Data/qkm5-nuaq>
      2. Traffic volume

<https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts-2014-2019-/ertz-hr4r>

* + 1. Citywide Mobility Survey

<https://data.cityofnewyork.us/Transportation/Citywide-Mobility-Survey-Household-Survey-2019/a5rk-jemi>

* + 1. Bicycle counts

<https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c>

<https://data.cityofnewyork.us/Transportation/Bicycle-Counters/smn3-rzf9>

* + 1. Traffic Bollards Tracking and Installations

<https://data.cityofnewyork.us/Transportation/Traffic-Bollards-Tracking-and-Installations/3f5t-9dqu>

* + 1. Taxi services

<https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Taxi+and+Limousine+Commission+%28TLC%29&page=3>

* + 1. Motor Vehicle Collisions - Vehicles

<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4>

* + 1. Motor Vehicle Collisions - Person

<https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu>

* + 1. Moving Violations

<https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-moving.page#citywide>

Some of the use cases of analysis on this dataset maybe

1. Transport companies/Taxi-cab service providers
2. Health service providers (ambulance/hospitals) - health data (and)
3. Insurance providers