**DSBA/MBAD 6211: Group Project**

**Learning Objectives:**

* To create an opportunity for students to work on real-world business analytics problems.
* To emphasize critical thinking, application of various analytics techniques, team work, and presentation skills.

**Grading:**

* Novelty/creativity: originality of thought; surprising way of approaching the data
* Integration of multiple analytics techniques to gain meaningful insights into the data
* Integration of multiple data sets to yield new insights
* Clarity of the oral presentation

**Project Deliverables:**

* Submit everything via Canvas:
  + Original data sets (only for groups that choose to work on their own data sets)
  + Data sources in addition to those provided by the professor, if any
  + A Word document
    - Project summary
    - Detailed analyses processes and steps
      * Enough details that allow others to replicate your results
    - Suggestions for additional data collection that might enhance future analyses
  + Presentation slides
    - A 15 minutes’ oral presentation to highlight selected analytics techniques and major findings

**Project Choices:**

* Option 1: An Empirical Investigation of a Local Airbnb Market
* Option 2: Data Science for Good: CareerVillage.org
* Option 3: Kaggle Machine Learning & Data Science Survey
* Option 4: Self-identified project
  + Subject to the professor’s approval
  + The dataset must
    - Be sizeable (# of records, # of variables)
    - Provide business/economic/social insights
    - Various data science competitions: <https://www.mltut.com/top-data-science-competitions/>

**Option 1: An Empirical Investigation of a Local Airbnb Market**

As part of the sharing economy platforms, Airbnb has expanded significantly since its founding in 2007 and become an alternative to traditional hotels. Please choose ***one city*** from the following list and analyze Airbnb data in the past 12 months in the selected local market.

Cities:

* Boston, MA
* Denver, CO
* Dublin, Ireland
* Edinburgh, UK
* New Orleans, LA
* Portland, OR
* Santa Clara County, CA

Time Range

* All available data from the website (the last 10-12 months)

Data:

* <http://insideairbnb.com/get-the-data.html>
* Please download monthly detailed listing, calendar, and review data.
* Additional data sources, such as census data and COVID-19 tracking data could be considered

Questions:

* Main objective: to gain insight into the selected city’s Airbnb market in the past year.
* Please propose and design a comprehensive list of matrices to understand the local Airbnb market. Examples of such measures are:
  + Market supply: the number of active listings
  + Market demand: occupancy rate, average monthly reviews per listing, etc.
  + Customer comments: topics, sentiments, etc.
* Do you observe any trend, seasonality, and neighborhood differences in your local market?
* Do you see any association between the Airbnb market activities and the Covid-19 situation in the local market?
  + Please note that customer review data goes beyond one year
* What factors affect Airbnb listings’ performance?
  + Please propose a reasonable indicator(s) to quantify individual listings’ performance.
    - For instance: average monthly occupancy rate, average monthly income, and/or average monthly reviews
  + Please identify suitable independent variables based on the available data sources.

**Option 2: Data Science for Good CareerVillage.org**

**(Adapted from a Kaggle Competition)**

CareerVillage.org is a nonprofit that crowdsources career advice for underserved youth. Founded in 2011 in four classrooms in New York City, the platform has now served career advice from 25,000 volunteer professionals to over 3.5M online learners. The platform uses a Q&A style similar to StackOverflow or Quora to provide students with answers to any question about any career.

The U.S. has almost 500 students for every guidance counselor. Underserved youth lack the network to find their career role models, making CareerVillage.org the only option for millions of young people in America and around the globe with nowhere else to turn.

To date, 25,000 volunteers have created profiles and opted in to receive emails when a career question is a good fit for them. This is where your skills come in. To help students get the advice they need, the team at CareerVillage.org needs to be able to send the right questions to the right volunteers. The notifications sent to volunteers seem to have the greatest impact on how many questions are answered.

Data:

* <https://www.kaggle.com/c/data-science-for-good-careervillage/data>

Questions:

* Please explore user profiles at CareerVillage
* Please analyze time series of community activities at CareerVillage.
* Please explore questions and answers posted at CareerVillage.
* What factors motivate users’ participation at CareerVillage?
  + For professionals
  + For students
  + Please propose multiple matrices to measure user activity levels, including both activity quantity and quality
* What factors affect a question’s likelihood to be answered?

**Option 3: Kaggle Machine Learning & Data Science Survey**

**(Adapted from a Kaggle Competition)**

Since 2018, Kaggle set out to conduct an annual industry-wide survey that presents a truly comprehensive view of the state of data science and machine learning. The results include raw numbers about who is working with data, what’s happening with machine learning in different industries, and the best ways for new data scientists to break into the field.

Please combine and dive deep into four years’ survey data, and understand various stories of data scientists from around the world

Data:

* 2018 Kaggle Machine Learning & Data Science Survey: <https://www.kaggle.com/kaggle/kaggle-survey-2018>
* 2019 Kaggle Machine Learning & Data Science Survey: <https://www.kaggle.com/c/kaggle-survey-2019/data>
* 2020 Kaggle Machine Learning & Data Science Survey: <https://www.kaggle.com/c/kaggle-survey-2020/data>
* 2021 Kaggle Machine Learning & Data Science Survey: <https://www.kaggle.com/c/kaggle-survey-2021/overview>

Questions:

* Please provide overviews to understand data scientists from around the world
* What changes, if there is any, have you observed from the 4 years’ survey data?
* Please explore regional differences in data scientist communities
* What is the value of higher education degrees in data science?
* What factors affect data scientists’ salaries?
* Please identify a specific group of data scientists, and tell a story about them
  + A “story” could be defined any number of ways through a combination of both narrative text and data exploration
  + You can define the sub-community in any way, such as female data scientists, data scientists from a specific region, or data scientists from a specific field etc.