Project Management Plan

Capstone Project

Addison, Lucas, Christian, Aaron

1) GitHub repository URL to start, dump files here (10/26/2021) - [1 hr] Aaron ✔
    a) Create repository Aaron ✔
    b) Permissions to other members Aaron ✔
    c) Start README Aaron ✔
    d) Include links to text files Aaron ✔
    e) Include links to visualizations Christian
    f) Include links to DataBricks stored in Jupyter Notebooks Christian
2) Make sure tasks are divided evenly: ✔
3) 5–10 exploratory questions(10/28/2021) [2 hr] ✔
    a) Create file Aaron ✔
    b) Permissions to other members Aaron ✔
    c) Write questions Aaron Christian ✔
4) Project management plan Aaron (10/29/2021)
    a) Create file Aaron, Lucas [1 hr] ✔
    b) Permissions to other members Aaron [1 hr]
    c) Update EODs [1 hr]
5) Get Data wrangled Addison, Lucas, Christian (11/05/2021)
    a) Start Data Bricks Addison, Lucas [1 hr] ✔
        i) Create data bricks for each producer
        ii) Create data bricks for each consumer
        iii) Create data brick for API call + data cleaning
    b) Find API [1 hr] Lucas, Aaron ✔
        i) Find API to call for mortality rate data
        ii) Find API to call for clean water data
        iii) Find API to call for BMI data
    c) Set up tables Lucas ✔ [2 hr]
        i) Create table for mortality rate data
        ii) Create table for clean water data
        iii) Create table for BMI data
    d) Clean data Addison, Lucas ✔ [2 hr]
        i) Clean mortality rate data
        ii) Clean the clean water data
        iii) Clean BMI data
        iv) Do not drop nulls, impute instead (noted below)
    e) Make producer for every table Addison, Lucas [2 hr] ✔
        i) Allows for parallel work

f) Import table for geographic regions Addison, Lucas [1 hr] ✔
g) Impute missing values using geographic regions [3 hr] ✔
h) Make producers Christian ✔ [1 hr]
    i) Do one producer per table Christian ✔
    ii) Use same mount point for each, but using different topics Christian ✔
i) Make consumers Addison, Lucas ✔ [1 hr]
    i) Send each consumed dataset to different csv file
j) Use Azure Data Factory to automate [2 hr] ✔
    i) Create pipelines Aaron
        (1) Mortality ✔
        (2) Water ✔
        (3) Underweight ✔
        (4) Tuberculosis
        (5) Malaria
    ii) Create consumers links Aaron ✔
    iii) Create producers links Aaron ✔
    iv) Add Wait time to consumers Addison ✔
    v) Use 'until' loop to trigger stop of consumers Addison ✔
        (1) Based off end of producers
    vi) Create Master Pipeline Lucas ✔
k) Read CSV files into SQL database [1hr] Addison ✔
    i) Write in separate folders for ease of separating. Addison
l) Construct region database (don't use producer or consumer) Addison, Lucas ✔
    i) Use the region tables imported beforehand
m) Get census data Addison, Lucas [1 hr] ✔
    i) Use CSVs downloaded
    ii) API
n) Clean census data Addison, Lucas [1 hr] ✔
    i) Examine specific states
    ii) Find comparable populations to other nations
o) ML Model (Linear?) [6 hr] ✔
    i) Time Series
    ii) Preprocess data
        (1) St scalar, etc
    iii) Run multiple
    iv) Other optimizations
        (1) Grid search/random search
p) Output to SQL database Addison ✔ [1 hr]
q) Connect to SQL database Addison ✔ [1 hr]

i) Verify with SQL queries

r) Automate data collection ✔ [3 hr]

i) Needs step e)

s) Use data in Power BI or Dash ✔

t) Create dashboard ✔

6) Napkin drawings Lucas, Christian, Aaron (10/29/2021) ✔

a) Dashboard ✔ [1 hr]

i) Get feedback from another group (10/29/2021) Christian ✔

ii) Working with Nicole (10/27/2021) Christian ✔

b) Visualizations [1 hr]

i) Get feedback from another group (10/29/2021) Aaron

ii) Working with Cory (10/27/2021) Aaron

7) Repeatable ETL Report (11/05/2021) [5 hr] ✔

a) Create ERD Christian ✔

b) Extraction ✔ [1 hr]

i) Identify Data Source(s) Aaron

ii) Show API queries Aaron

iii) Show web scraping Aaron

c) Transformation ✔ [2 hr]

i) Show Column renames Aaron

ii) Show Column drops Aaron

iii) Show Data imputation Aaron

iv) Show JSON transformation Aaron

v) Show Kafka Producer(s) sending to Topic(s) Aaron

vi) Show Kafka Consumer(s) reading from Topic(s) Aaron

vii) Show Kafka Consumer(s) transforming to csv Aaron

d) Loading ✔ [1 hr]

i) Show Kafka Consumer(s) saving csv to Data Lake Aaron

ii) Show csvReader writing data from Data Lake csvs to SQL database Aaron

e) Editing Christian ✔

i) Grammar Christian

ii) Typos Christian

iii) Structural editing Christian

8) Dashboard with Power BI or Dash (11/05/2021) [3 hr] ✔

a) Build prototype to test usability of Power BI Aaron ✔

b) Confirm visualizations Aaron ✔

c) Confirm overall design Aaron ✔

d) Confirm has clear messaging Aaron ✔

e) Create first draft Aaron ✔

f) Get Feedback Aaron ✔

g) Create Second draft (final?) Aaron
    i) Tweak as needed Aaron

9) Project Executive Summary (11/10/2021) [5 hr] ✔
    a) Create file Aaron ✔
    b) Permissions to group members Aaron ✔
    c) Introduction Aaron ✔
    d) Research Aaron , Lucas ✔
    e) Conclusion Lucas ✔
10) Presentation Slides (11/10/2021) [2 hr] ✔
    a) Create slides document Aaron ✔
    b) Permissions to group members Aaron ✔
    c) Decide on theme/colors/font/visuals Christian ✔
    d) Intro Christian ✔
    e) Initial Questions Christian ✔
    f) Research process Christian ✔
    g) Machine Learning Christian ✔
    h) Recommendations Christian ✔
    i) Conclusion Christian ✔
    j) Works Cited Christian ✔
11) Presentation Dry Run (11/10/2021) [2 hr] (ASSIGNMENTS TBD)
    a) Introduction
    b) Initial questions
    c) Research process
    d) Machine Learning
    e) Recommendations
    f) Conclusion
    g) Tweak/edit/fix hangups (WHOLE GROUP)

DELIVERABLES to keep in mind:

- GitHub URL ✔
- Exploratory Questions ✔
- Project Management Plan ✔
- Napkin Drawings/Feedback -- Visualizations ✔
- Napkin Drawings/Feedback -- Dashboard ✔

- Repeatable ETL Report ✔
    - Requires project management plan to be finalized
- Dashboard ✔
    - Requires ETL to be finished
- Project Executive Summary ✔
- Presentation Slides
    - Requires napkin drawings
    - Requires dashboard
- README
- Organization of Git
    - Requires ETL completion
    - Requires other exercises