Variable: Medical Costs
Adults in the US

Link to data:
https://www.kaggle.com/mirichoi0218/insurance

Columns:
Age: age of primary beneficiary
Sex: insurance contractor gender, female, male
Bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
Children: Number of children covered by health insurance / Number of dependents
Smoker: Smoking
Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Charges: Individual medical costs billed by health insurance

Questions:
-How much do health charges increase over the span of someone's life?
-How do smokers' versus non-smokers' health charges increase over the span of their life?
-Who has higher health charges, men or women?
-How does BMI in different brackets (underweight, average, obese etc) impact health costs for women with varied numbers of children?
-How does the number of children impact health costs for men versus women?
-Do "younger" mothers have the same medical costs as "older" mothers?
-Which variables have the largest differences between regions?
-Which regions have the highest health charges?
-What indicator has the largest impact on health costs for people in different age brackets?
-Does a high BMI impact male health charges to the same degree that a high BMI impacts women's health charges?

Predictive analytics / machine learning
Use all components to predict cost

https://github.com/Esavoye/Introduction-To-Data-Science/blob/master/Capstone-Project/Capstone%20Proposal.Rmd

Approach and deliverables will be the same as the model.

**POST PROJECTS ON PLACEHOLDER ON THE ACCOUNT*
-in agenda write which projects I've completed

Send emails with more information. Agenda is quick bullet list.

-send email if get stuck with the project.

2. Variable: Sales Price
DC Residential Properties

Link:
https://www.kaggle.com/christophercorrea/dc-residential-properties

Columns:
Full Bathrooms (number)
Half Bathrooms (number)
Heat (type)
AC (yes/no)
Units (number)
Rooms (number)
Bedrooms (number)
AYB-earliest time the main portion of the building was build
YR_RMDL-Year structure was remodeled
EYB-Year and improvement was build more recent than actual year built
Stories (number of floors)
Sales Date
Price
Qualified (yes/no)
Sale Number
GBA-gross building area in square feet
Building Number Property
Style
Structure

-How have DC residential sales prices changed over time?
-How have DC residential sales prices changed over time for different styles of buildings?
-How does gross building area impact cost of properties with different numbers of units?
-How much of an impact does the heat and ac type impact the sales price?
-How much of an impact does the Year Remodeled impact the price versus AYB (earliest time the main portion of the building was built).
-What impacts the sales price the most, the number of rooms versus number of bedrooms?
-What impacts the sales price the most, the number of full bathrooms (examine values over 1), or the number of half-bathrooms?
-When looking at the ratio of stories to the number of units, which ratios have the highest sales price?
-When looking at the ratio of gross building area to the number units, how does that ratio impact price?
-What impact does structure have on sales price?

I'm interested in the Geographical location/segregation and its relationship to sales price.
Can I combine this data set with another one to ask the above questions in relationship to location/race?
How would I match the data???

Geographical/racial data:
https://www.kaggle.com/christophercorrea/dc-residential-properties#raw_address_points.csv

3. Variable: Number of views
Ted Talks

Link:
https://www.kaggle.com/rounakbanik/ted-talks

These datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017.

Columns:
-First-level comments (number)
-Description (character)
-Duration of talk in seconds
-The TED/TEDx event where the talk took place
-Film date (Unix timestamp)
-Languages available (number)
-Main speaker (name)
-Title of the Ted Talk
-Number of speakers in the talk
-Published date (Unix timestamp)
-Ratings (stringified dictionary, express emotional impact of talk)
-Related talks (list of dictionaries of recommended talks)
-Speaker occupation
-Tags (themes associated with the talk)
-Title of the talk
-URL link
-Number of views

Questions:
-How have the total number of views changed over time?
-Is there a correlation between the number of first-level comments and number of views?
-Which tags correlate with the highest number of views?
*there are multiple tags for each talk, so I'm unsure about how to analyze this accurately.
-Does the number of languages available impact the number of views? Has the number of languages available over time changed?
-Which ratings correlate with the highest number of views?
*there are multiple ratings for each talk, so I'm unsure about how to analyze this accurately.
-Which TED speakers get the most views? Is there anything that those speakers have in common?
-Which "related talks" are suggested the most? Do those talks have more views than "related talks" that are suggested less?
-Which TED events had the talks with the greatest number of views?
-Is there a correlation between the duration of the talk and number of views?