Team 1: Tyler Kutz,Kavita Jain, Pallabi Chatterjee, Ajinkya Koshti, Ludmila Scott
10/1/19
Data Analytics Project - Checkpoint 1

1. ***How does your chosen topic and identified data and supporting material satisfy each one of the 5 criteria below. Please see the explanation provided above for each criteria in "Five Criteria for Appropriate Data " guideline above.***

    a. ***Importance****:*
    Cancer is a public health issue. Investigating behavioral patterns that can influence cancer risk or cancer recurrence is important. Since there is no definitive cure for cancer, it becomes crucial to work towards early detection measures, treatments, and follow-up methods that are proven to save lives.

    b. ***Availability****:*
    The cancer data is available through the SEER website. SEER is an authoritative source for cancer statistics in the United States.

    c. ***Documentation:***
    SEER provides detailed documentation and glossary of statistical terms in the website. It has also made videos available to us that helps us navigate with the data in use.

    d. ***Support:***
    There is a "contact us" area on the website. The website is supported by Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS)

    e. ***Size:***
    The 2000-2016 colon/rectal cancer dataset has 939,119 records, 35 columns and has a file size of 200 MB.

2. ***Describe your data properties, including the following, as much as possible.***
    a. ***Data format (tabular, database or file format, etc.)***
    The data was previously in the text format which was converted into a csv file. The file has coded domains, so that may need to be changed later to facilitate analysis.

    b. ***Data tables (how many, their content/organization, etc.)***
    There was one data table for all of the colon cancer data. There is a data dictionary that might need to be inputted into a single table at some point.

c. ***Data columns (most important ones, etc.)***
   Race
   Sex
   Age
   Year of birth
   Month of diagnosis
   Year of diagnosis
   Type of reporting source
   Grade
   First indicator
   State-county
   Primary Site
   Laterality
   Histologic Type ICD-O-3
   Behavior Code ICD-O-3
   Diagnostic Confirmation
   Type of Reporting Source
   IHS Link (Native American Heritage)

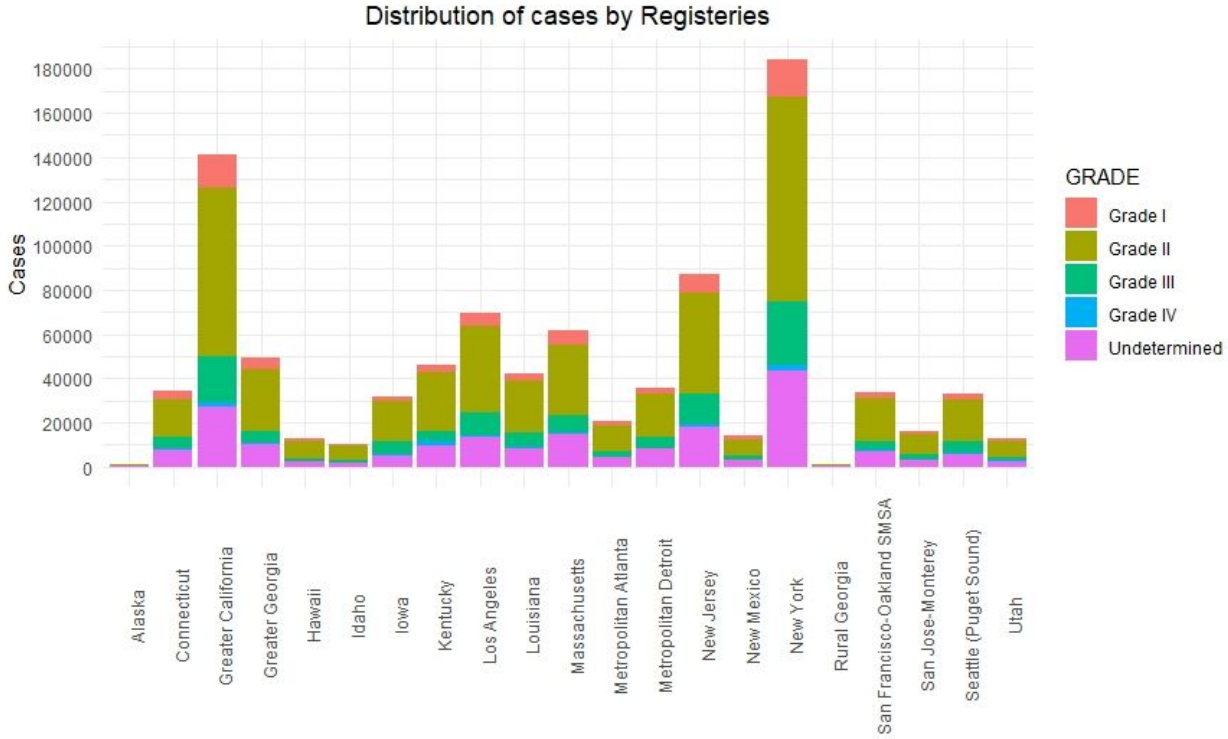
d. ***Data rows (unit of observation, count, etc.)***
   The data rows describe individual occurrences of cancer in a patient. Each row is a different case of cancer.

3. ***Describe your data variables. Please use distribution statistics (mean, median, mode, percent missing, etc.) and distribution charts.***

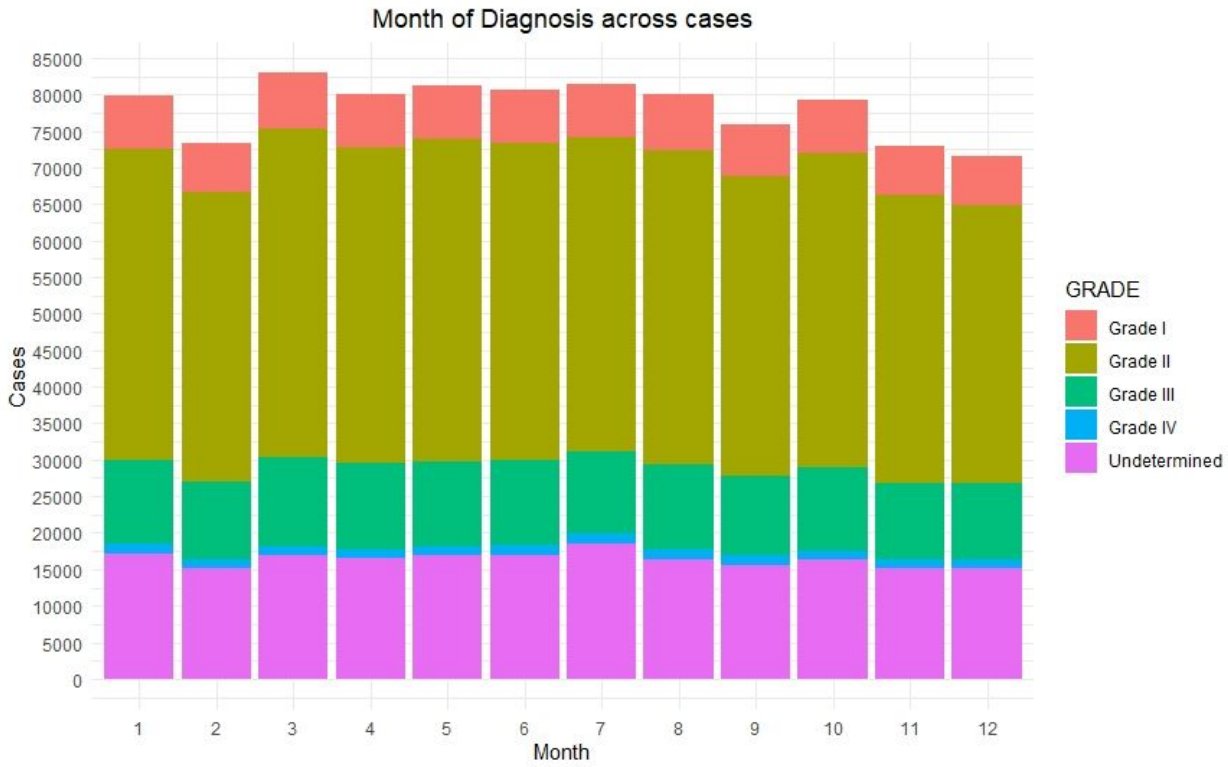
Variables	*Mean*	*Std Dev*	*Minimum*	*Maximum*
Age of Diagnosis	***67.56***	***14.23***	***0***	***116***
Year of Birth	***1939***	***15.266***	***1893***	***2011***
Year of Diagnosis	***2008***	***4.91***	***2000***	***2016***

a. Categorical variables (nominal or ordinal):

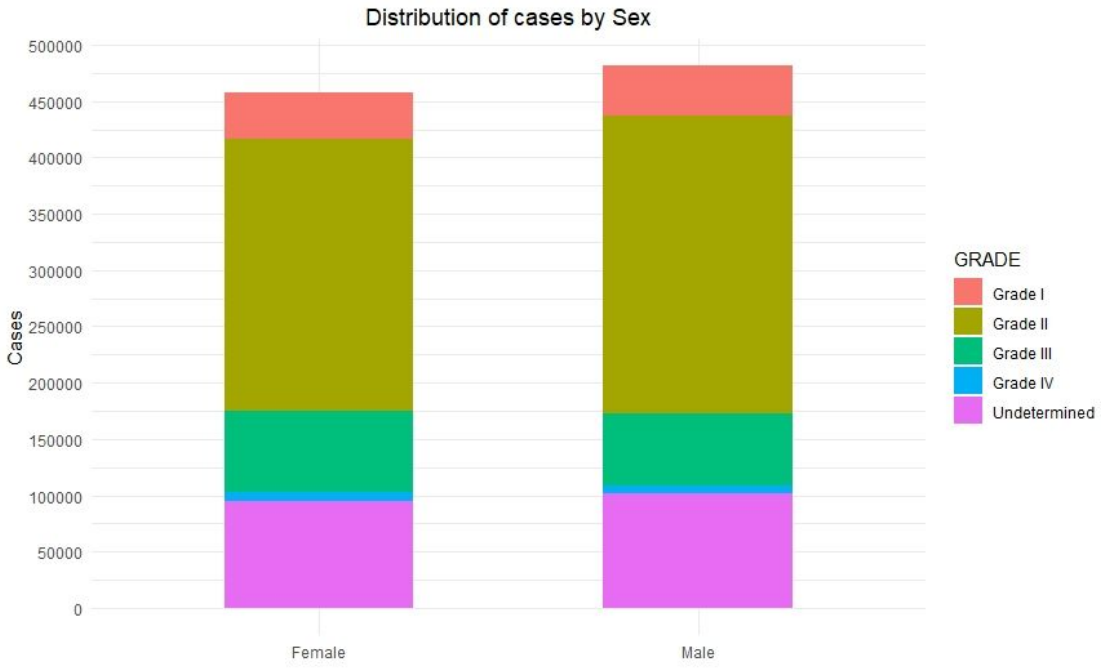
Registry ID



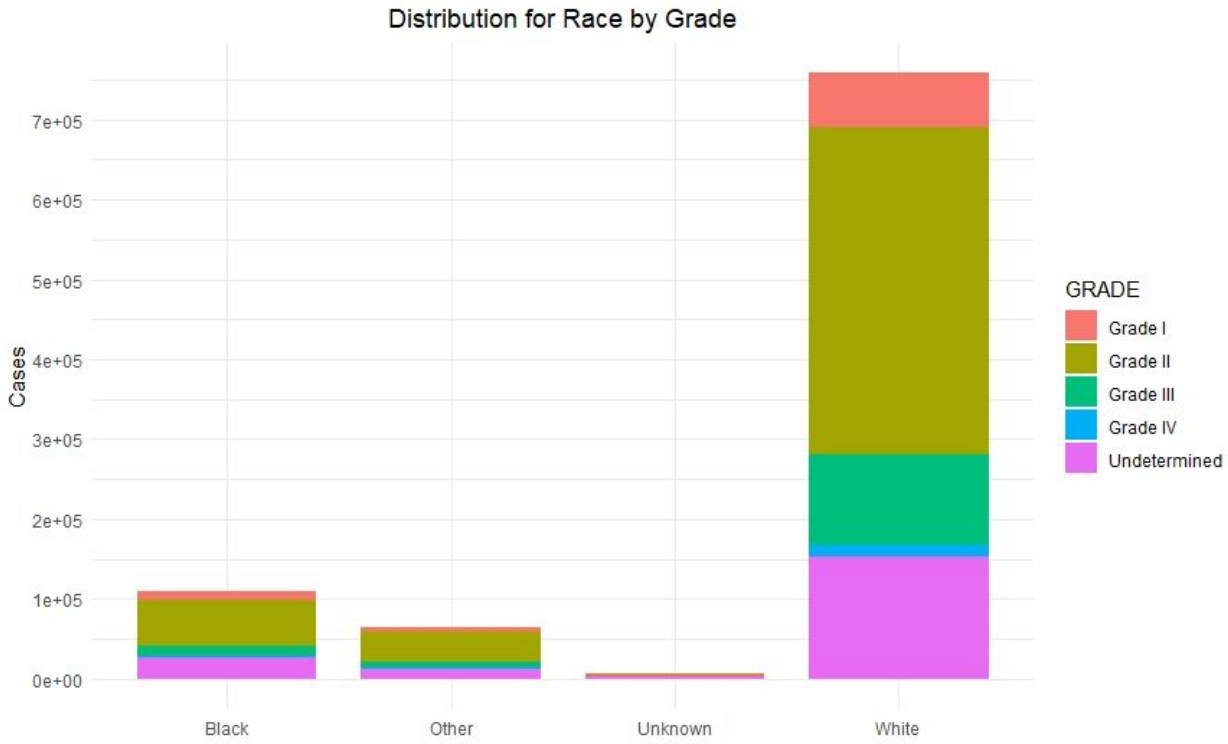
Month of diagnosis (ordinal)



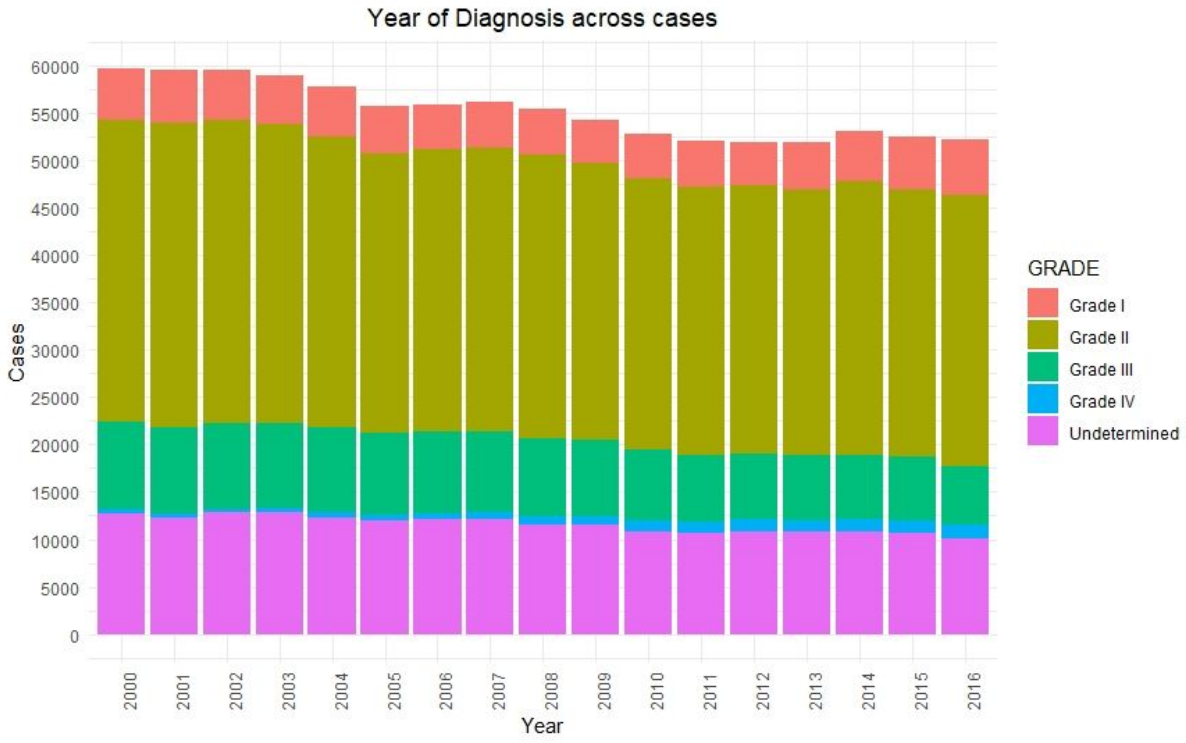
Sex (nominal)



Race (nominal)

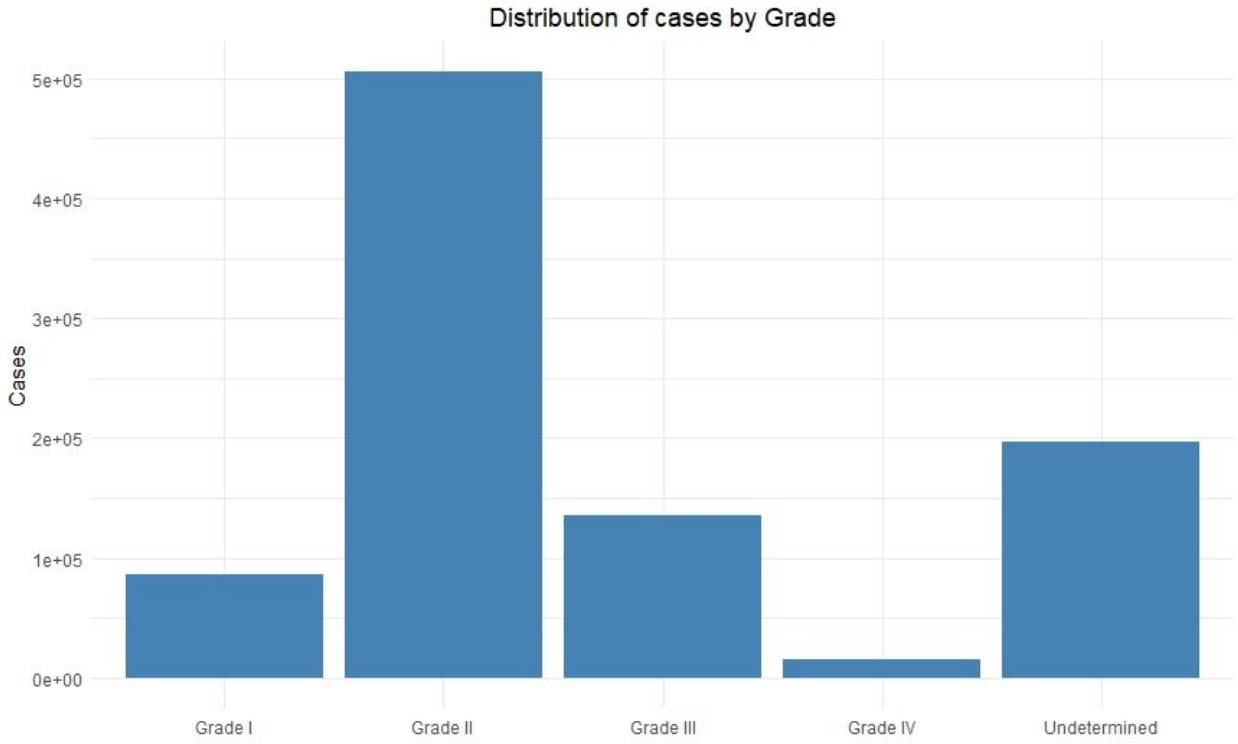


Year of diagnosis (ordinal)

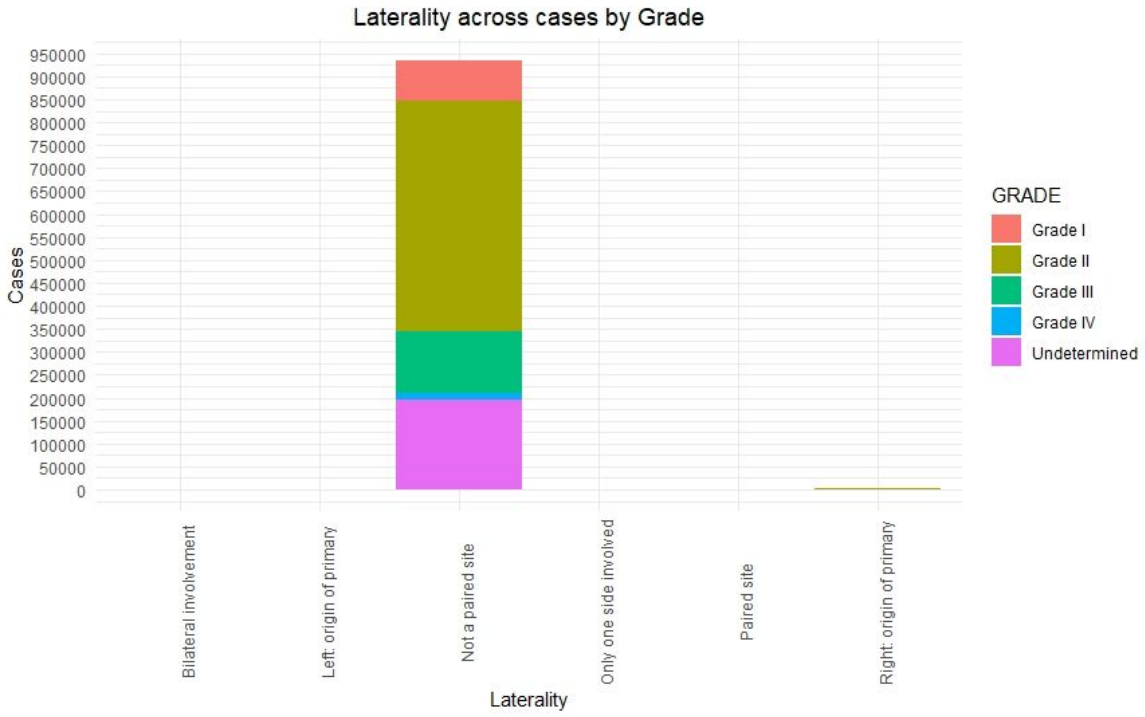


Type of reporting source (nominal)

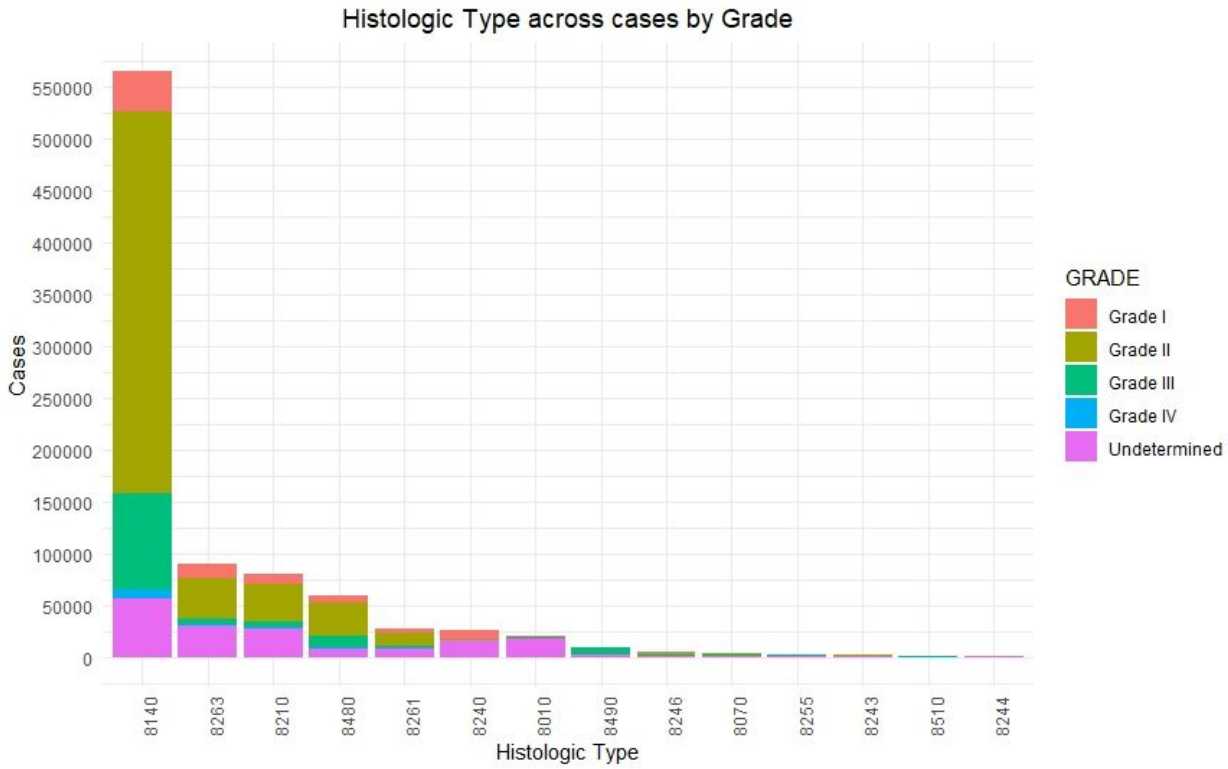
Grade (ordinal)



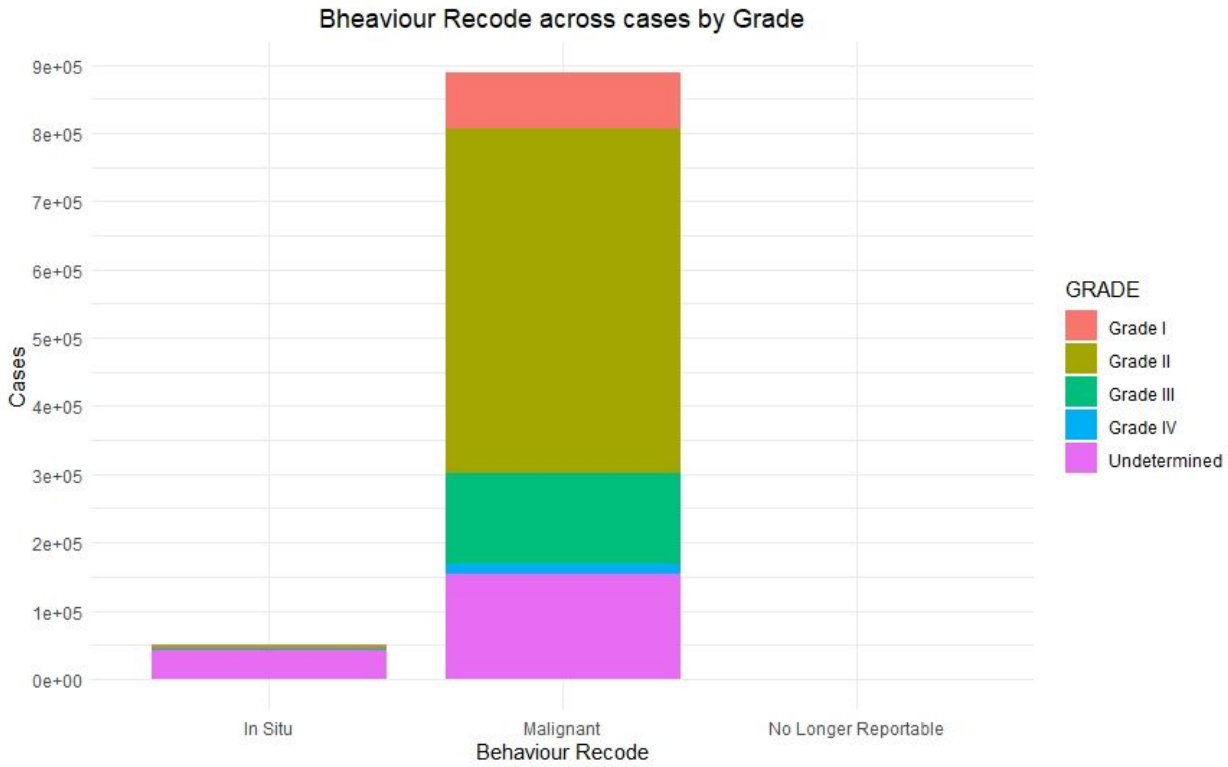
Laterality (nominal)



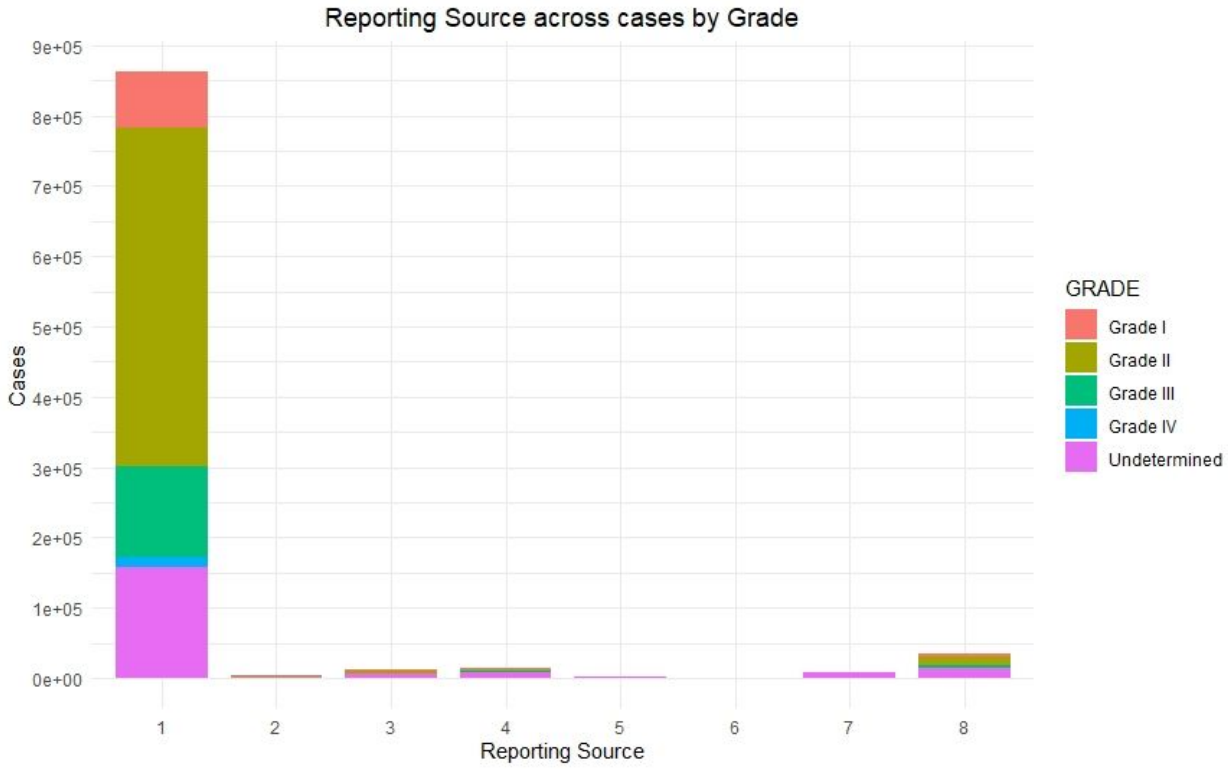
Histologic Type ICD-O-3 (nominal)



Behavior Code ICD-O-3 (nominal)



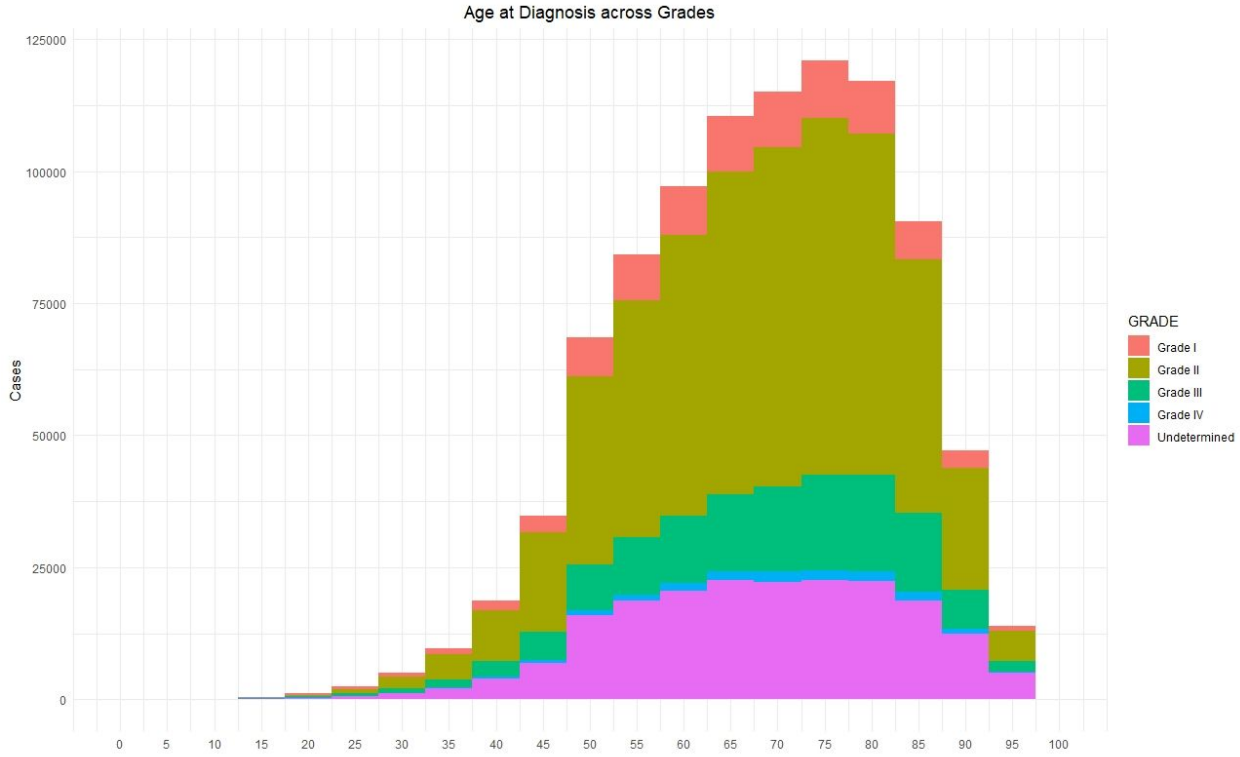
Type of Reporting Source (nominal)



Patient ID Number
Year of birth (ordinal)
IHS Link (binary)
Diagnostic Confirmation (nominal)
Primary Site (nominal)
First indicator (nominal)

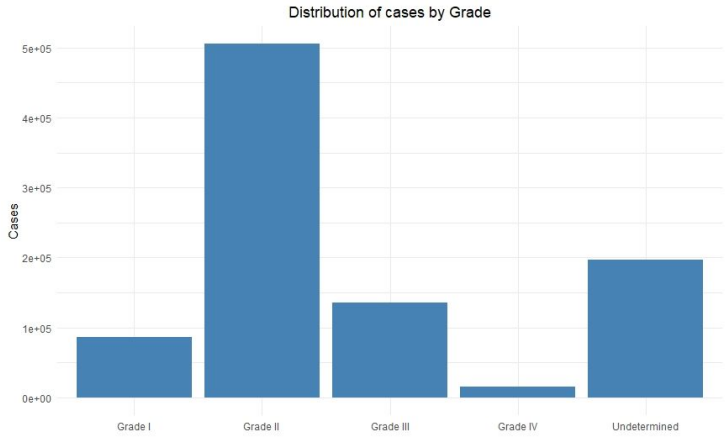
b. Numerical variables (binary or interval)

Age (interval)

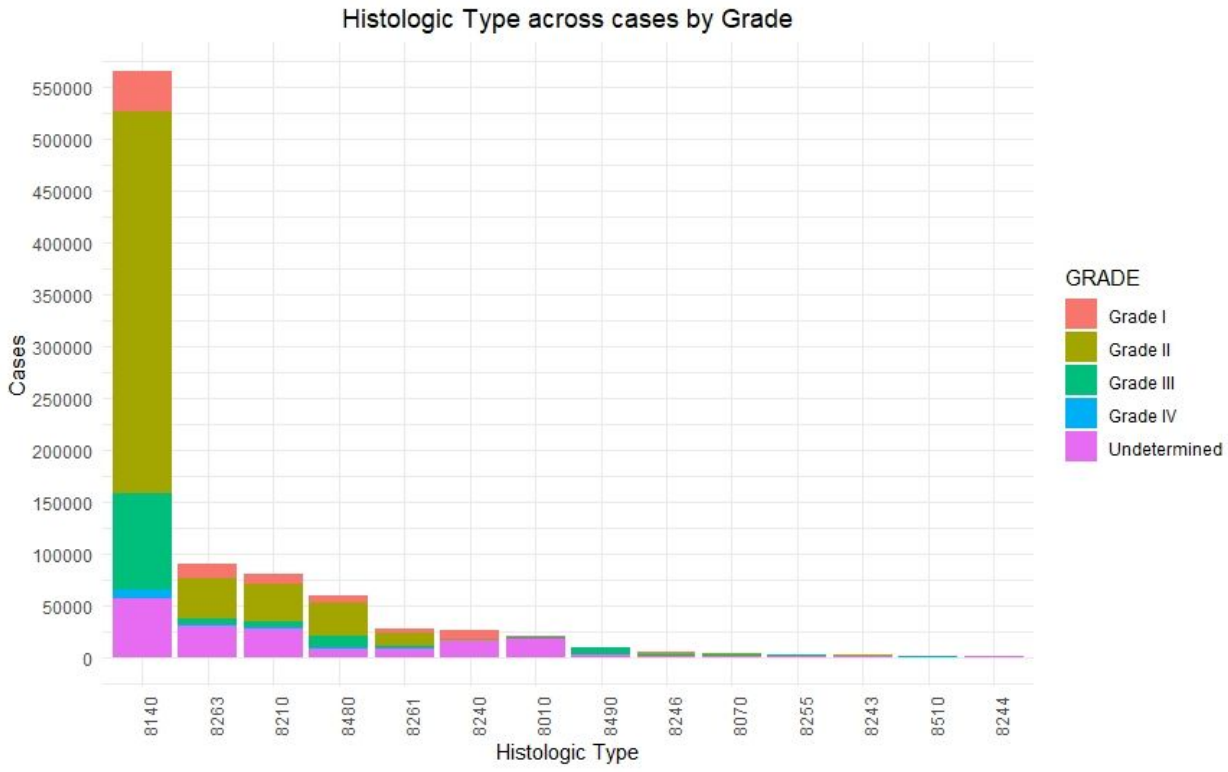


c. Potential target variable:

Grade (whether the cancer is of Grade I, II, III or IV)



Histological Type



4. ***Propose potential questions you will answer with or insights you will gain from your data analytics.***

    a) How are histological type, primary site and biological type related to grade of colon cancer and thus the progression of cancer?
    b) Do race, age, geography, and other demographics affect the grade of colon cancer and thus the progression of cancer?
    c) Do race, age, and other demographics affect the histological type? Are certain types of people more likely to have cancer develop from a certain tissue?
    d) Are there certain demographics or geographic areas that will lead to higher incidence rates?