Mini Data-Analysis Deliverable 1

Welcome to your (maybe) first-ever data analysis project!

And hopefully the first of many. Let's get started:

1. Install the `datateachr` package by typing the following into your **R terminal**:

```
install.packages("devtools")
devtools::install_github("UBC-MDS/datateachr")
```

2. Load the packages below.

```
library(datateachr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

3. Make a repository in the https://github.com/stat545ubc-2024 Organization. You can do this by following the steps found on canvas in the entry called MDA: Create a repository. One completed, your repository should automatically be listed as part of the stat545ubc-2024 Organization.

Instructions

For Both Milestones

- Each milestone has explicit tasks. Tasks that are more challenging will often be allocated more points.
- Each milestone will be also graded for reproducibility, cleanliness, and coherence of the overall Github submission.
- While the two milestones will be submitted as independent deliverables, the analysis itself is a continuum - think of it as two chapters to a story. Each chapter, or in this case, portion of your analysis, should be easily followed through by someone unfamiliar with the content. Here is a good resource for what constitutes "good code". Learning good coding practices early in your career will save you hassle later on!
- The milestones will be equally weighted.

For Milestone 1

To complete this milestone, edit this very `.Rmd` file directly. Fill in the sections that are tagged with `<!--- start your work below --->`.

To submit this milestone, make sure to knit this `.Rmd` file to an `.md` file by changing the YAML output settings from `output: html_document` to `output: github_document`. Commit and push all of your work to the mini-analysis GitHub repository you made earlier, and tag a release on GitHub. Then, submit a link to your tagged release on canvas.

Points: This milestone is worth 36 points: 30 for your analysis, and 6 for overall reproducibility, cleanliness, and coherence of the Github submission.

Learning Objectives

By the end of this milestone, you should:

- Become familiar with your dataset of choosing
- Select 4 questions that you would like to answer with your data
- Generate a reproducible and clear report using R Markdown
- Become familiar with manipulating and summarizing your data in tibbles using `dplyr`, with a research question in mind.

Task 1: Choose your favorite dataset

The `datateachr` package by Hayley Boyce and Jordan Bourak currently composed of 7 semi-tidy datasets for educational purposes. Here is a brief description of each dataset:

- *apt_buildings*: Acquired courtesy of The City of Toronto's Open Data Portal. It currently has 3455 rows and 37 columns.
- *building_permits*: Acquired courtesy of The City of Vancouver's Open Data Portal. It currently has 20680 rows and 14 columns.
- *cancer_sample*: Acquired courtesy of UCI Machine Learning Repository. It currently has 569 rows and 32 columns.
- *flow_sample*: Acquired courtesy of The Government of Canada's Historical Hydrometric Database. It currently has 218 rows and 7 columns.
- *parking_meters*: Acquired courtesy of The City of Vancouver's Open Data Portal. It currently has 10032 rows and 22 columns.
- *steam_games*: Acquired courtesy of Kaggle. It currently has 40833 rows and 21 columns.
- *vancouver_trees*: Acquired courtesy of The City of Vancouver's Open Data Portal. It currently has 146611 rows and 20 columns.

Things to keep in mind

- We hope that this project will serve as practice for carrying our your own *independent* data analysis. Remember to comment your code, be explicit about what you are doing, and write notes in this markdown document when you feel that context is required. As you advance in the project, prompts and hints to do this will be diminished - it'll be up to you!

- Before choosing a dataset, you should always keep in mind **your goal**, or in other ways, *what you wish to achieve with this data*. This mini data-analysis project focuses on *data wrangling*, *tidying*, and *visualization*. In short, it's a way for you to get your feet wet with exploring data on your own.

And that is exactly the first thing that you will do!

1.1 **(1 point)** Out of the 7 datasets available in the `datateachr` package, choose **4** that appeal to you based on their description. Write your choices below:

Note: We encourage you to use the ones in the `datateachr` package, but if you have a dataset that you'd really like to use, you can include it here. But, please check with a member of the teaching team to see whether the dataset is of appropriate complexity. Also, include a **brief** description of the dataset here to help the teaching team understand your data.

1: apt_buildings

2: cancer_sample

3: flow_sample

4: steam_games

1.2 **(6 points)** One way to narrowing down your selection is to *explore* the datasets. Use your knowledge of dplyr to find out at least *3* attributes about each of these datasets (an attribute is something such as number of rows, variables, class type...). The goal here is to have an idea of *what the data looks like*.

Hint: This is one of those times when you should think about the cleanliness of your analysis. I added a single code chunk for you below, but do you want to use more than one? Would you like to write more comments outside of the code chunk?

```
### EXPLORE HERE ###
# Use glimpse to visualize the information of each dataset
# Split datasets by names for clarity

cat("Dataset: apt_buildings\n")
```

```
## Dataset: apt_buildings
```

```
glimpse(apt_buildings)
```

```
## Rows: 3,455
## Columns: 37
## $ id                                <dbl> 10359, 10360, 10361, 10362, 10363, 10~
## $ air_conditioning                  <chr> "NONE", "NONE", "NONE", "NONE", "NONE~
## $ amenities                         <chr> "Outdoor rec facilities", "Outdoor po~
## $ balconies                         <chr> "YES", "YES", "YES", "YES", "NO", "NO~
## $ barrier_free_accessibilty_entr    <chr> "YES", "NO", "NO", "YES", "NO", "NO",~
## $ bike_parking                      <chr> "0 indoor parking spots and 10 outdoo~
## $ exterior_fire_escape              <chr> "NO", "NO", "NO", "YES", "NO", NA, "N~
## $ fire_alarm                        <chr> "YES", "YES", "YES", "YES", "YES", "Y~
## $ garbage_chutes                    <chr> "YES", "YES", "NO", "NO", "NO", "NO",~
## $ heating_type                      <chr> "HOT WATER", "HOT WATER", "HOT WATER"~
## $ intercom                          <chr> "YES", "YES", "YES", "YES", "YES", "Y~
## $ laundry_room                      <chr> "YES", "YES", "YES", "YES", "YES", "Y~
## $ locker_or_storage_room            <chr> "NO", "YES", "YES", "YES", "NO", "YES~
## $ no_of_elevators                   <dbl> 3, 3, 0, 1, 0, 0, 0, 2, 4, 2, 0, 2, 2~
```

```
## $ parking_type                     <chr> "Underground Garage , Garage accessib~
## $ pets_allowed                     <chr> "YES", "YES", "YES", "YES", "YES", "Y~
## $ prop_management_company_name     <chr> NA, "SCHICKEDANZ BROS. PROPERTIES", N~
## $ property_type                    <chr> "PRIVATE", "PRIVATE", "PRIVATE", "PRI~
## $ rsn                              <dbl> 4154812, 4154815, 4155295, 4155309, 4~
## $ separate_gas_meters              <chr> "NO", "NO", "NO", "NO", "NO", "NO", "~
## $ separate_hydro_meters            <chr> "YES", "YES", "YES", "YES", "YES", "Y~
## $ separate_water_meters            <chr> "NO", "NO", "NO", "NO", "NO", "NO", "~
## $ site_address                     <chr> "65  FOREST MANOR RD", "70  CLIPPER R~
## $ sprinkler_system                 <chr> "YES", "YES", "NO", "YES", "NO", "NO"~
## $ visitor_parking                  <chr> "PAID", "FREE", "UNAVAILABLE", "UNAVA~
## $ ward                             <chr> "17", "17", "03", "03", "02", "02", "~
## $ window_type                      <chr> "DOUBLE PANE", "DOUBLE PANE", "DOUBLE~
## $ year_built                       <dbl> 1967, 1970, 1927, 1959, 1943, 1952, 1~
## $ year_registered                  <dbl> 2017, 2017, 2017, 2017, 2017, NA, 201~
## $ no_of_storeys                    <dbl> 17, 14, 4, 5, 4, 4, 4, 7, 32, 4, 4, 7~
## $ emergency_power                  <chr> "NO", "YES", "NO", "NO", "NO", "NO", ~
## $ `non-smoking_building`           <chr> "YES", "NO", "YES", "YES", "YES", "NO~
## $ no_of_units                      <dbl> 218, 206, 34, 42, 25, 34, 14, 105, 57~
## $ no_of_accessible_parking_spaces  <dbl> 8, 10, 20, 42, 12, 0, 5, 1, 1, 6, 12,~
## $ facilities_available             <chr> "Recycling bins", "Green Bin / Organi~
## $ cooling_room                     <chr> "NO", "NO", "NO", "NO", "NO", "NO", "~
## $ no_barrier_free_accessible_units <dbl> 2, 0, 0, 42, 0, NA, 14, 0, 0, 1, 25, ~
```

```
cat("\nDataset: cancer_sample\n")
```

```
##
## Dataset: cancer_sample
```

```
glimpse(cancer_sample)
```

```
## Rows: 569
## Columns: 32
## $ ID                      <dbl> 842302, 842517, 84300903, 84348301, 84358402, ~
## $ diagnosis               <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ radius_mean             <dbl> 17.990, 20.570, 19.690, 11.420, 20.290, 12.450~
## $ texture_mean            <dbl> 10.38, 17.77, 21.25, 20.38, 14.34, 15.70, 19.9~
## $ perimeter_mean          <dbl> 122.80, 132.90, 130.00, 77.58, 135.10, 82.57, ~
## $ area_mean               <dbl> 1001.0, 1326.0, 1203.0, 386.1, 1297.0, 477.1, ~
## $ smoothness_mean         <dbl> 0.11840, 0.08474, 0.10960, 0.14250, 0.10030, 0~
## $ compactness_mean        <dbl> 0.27760, 0.07864, 0.15990, 0.28390, 0.13280, 0~
## $ concavity_mean          <dbl> 0.30010, 0.08690, 0.19740, 0.24140, 0.19800, 0~
## $ concave_points_mean     <dbl> 0.14710, 0.07017, 0.12790, 0.10520, 0.10430, 0~
## $ symmetry_mean           <dbl> 0.2419, 0.1812, 0.2069, 0.2597, 0.1809, 0.2087~
## $ fractal_dimension_mean  <dbl> 0.07871, 0.05667, 0.05999, 0.09744, 0.05883, 0~
## $ radius_se               <dbl> 1.0950, 0.5435, 0.7456, 0.4956, 0.7572, 0.3345~
## $ texture_se              <dbl> 0.9053, 0.7339, 0.7869, 1.1560, 0.7813, 0.8902~
## $ perimeter_se            <dbl> 8.589, 3.398, 4.585, 3.445, 5.438, 2.217, 3.18~
## $ area_se                 <dbl> 153.40, 74.08, 94.03, 27.23, 94.44, 27.19, 53.~
## $ smoothness_se           <dbl> 0.006399, 0.005225, 0.006150, 0.009110, 0.0114~
## $ compactness_se          <dbl> 0.049040, 0.013080, 0.040060, 0.074580, 0.0246~
## $ concavity_se            <dbl> 0.05373, 0.01860, 0.03832, 0.05661, 0.05688, 0~
## $ concave_points_se       <dbl> 0.015870, 0.013400, 0.020580, 0.018670, 0.0188~
```

```
## $ symmetry_se             <dbl> 0.03003, 0.01389, 0.02250, 0.05963, 0.01756, 0~
## $ fractal_dimension_se    <dbl> 0.006193, 0.003532, 0.004571, 0.009208, 0.0051~
## $ radius_worst            <dbl> 25.38, 24.99, 23.57, 14.91, 22.54, 15.47, 22.8~
## $ texture_worst           <dbl> 17.33, 23.41, 25.53, 26.50, 16.67, 23.75, 27.6~
## $ perimeter_worst         <dbl> 184.60, 158.80, 152.50, 98.87, 152.20, 103.40,~
## $ area_worst              <dbl> 2019.0, 1956.0, 1709.0, 567.7, 1575.0, 741.6, ~
## $ smoothness_worst        <dbl> 0.1622, 0.1238, 0.1444, 0.2098, 0.1374, 0.1791~
## $ compactness_worst       <dbl> 0.6656, 0.1866, 0.4245, 0.8663, 0.2050, 0.5249~
## $ concavity_worst         <dbl> 0.71190, 0.24160, 0.45040, 0.68690, 0.40000, 0~
## $ concave_points_worst    <dbl> 0.26540, 0.18600, 0.24300, 0.25750, 0.16250, 0~
## $ symmetry_worst          <dbl> 0.4601, 0.2750, 0.3613, 0.6638, 0.2364, 0.3985~
## $ fractal_dimension_worst <dbl> 0.11890, 0.08902, 0.08758, 0.17300, 0.07678, 0~
```

```
cat("\nDataset: flow_sample\n")
```

```
##
## Dataset: flow_sample
```

```
glimpse(flow_sample)
```

```
## Rows: 218
## Columns: 7
## $ station_id   <chr> "05BB001", "05BB001", "05BB001", "05BB001", "05BB001", "0~
## $ year         <dbl> 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 191~
## $ extreme_type <chr> "maximum", "maximum", "maximum", "maximum", "maximum", "m~
## $ month        <dbl> 7, 6, 6, 8, 6, 6, 6, 6, 6, 6, 6, 7, 6, 6, 6, 7, 5, 7, 6, ~
## $ day          <dbl> 7, 12, 14, 25, 11, 18, 27, 20, 17, 15, 22, 3, 9, 5, 14, 5~
## $ flow         <dbl> 314, 230, 264, 174, 232, 214, 236, 309, 174, 345, 185, 24~
## $ sym          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
cat("\nDataset: steam_games\n")
```

```
##
## Dataset: steam_games
```

```
glimpse(steam_games)
```

```
## Rows: 40,833
## Columns: 21
## $ id                      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14~
## $ url                     <chr> "https://store.steampowered.com/app/379720/DO~
## $ types                   <chr> "app", "app", "app", "app", "app", "bundle", ~
## $ name                    <chr> "DOOM", "PLAYERUNKNOWN'S BATTLEGROUNDS", "BAT~
## $ desc_snippet            <chr> "Now includes all three premium DLC packs (Un~
## $ recent_reviews          <chr> "Very Positive,(554),- 89% of the 554 user re~
## $ all_reviews             <chr> "Very Positive,(42,550),- 92% of the 42,550 u~
## $ release_date            <chr> "May 12, 2016", "Dec 21, 2017", "Apr 24, 2018~
## $ developer               <chr> "id Software", "PUBG Corporation", "Harebrain~
## $ publisher               <chr> "Bethesda Softworks,Bethesda Softworks", "PUB~
## $ popular_tags            <chr> "FPS,Gore,Action,Demons,Shooter,First-Person,~
## $ game_details            <chr> "Single-player,Multi-player,Co-op,Steam Achie~
```

```
## $ languages                <chr> "English,French,Italian,German,Spanish - Spai~
## $ achievements             <dbl> 54, 37, 128, NA, NA, NA, 51, 55, 34, 43, 72, ~
## $ genre                    <chr> "Action", "Action,Adventure,Massively Multipl~
## $ game_description         <chr> "About This Game Developed by id software, th~
## $ mature_content           <chr> NA, "Mature Content Description  The develope~
## $ minimum_requirements     <chr> "Minimum:,OS:,Windows 7/8.1/10 (64-bit versio~
## $ recommended_requirements <chr> "Recommended:,OS:,Windows 7/8.1/10 (64-bit ve~
## $ original_price           <dbl> 19.99, 29.99, 39.99, 44.99, 0.00, NA, 59.99, ~
## $ discount_price           <dbl> 14.99, NA, NA, NA, NA, 35.18, 70.42, 17.58, N~
```

1.3 **(1 point)** Now that you've explored the 4 datasets that you were initially most interested in, let's narrow it down to 1. What lead you to choose this one? Briefly explain your choice below.

I am most interested in cancer_sample. From the examples, the dataset contains different categories of data, including string and number, and dataset has no missing values, showing better data quality.

1.4 **(2 points)** Time for a final decision! Going back to the beginning, it's important to have an *end goal* in mind. For example, if I had chosen the `titanic` dataset for my project, I might've wanted to explore the relationship between survival and other variables. Try to think of 1 research question that you would want to answer with your dataset. Note it down below.

I want to explore the relationship between cancer diagnosis and other variables, for example, what combinations of tumor characteristics best distinguish between malignant and benign breast cancer diagnoses, and how do these features interact?

Important note

Read Tasks 2 and 3 *fully* before starting to complete either of them. Probably also a good point to grab a coffee to get ready for the fun part!

This project is semi-guided, but meant to be *independent*. For this reason, you will complete tasks 2 and 3 below (under the **START HERE** mark) as if you were writing your own exploratory data analysis report, and this guidance never existed! Feel free to add a brief introduction section to your project, format the document with markdown syntax as you deem appropriate, and structure the analysis as you deem appropriate. If you feel lost, you can find a sample data analysis here to have a better idea. However, bear in mind that it is **just an example** and you will not be required to have that level of complexity in your project.

Task 2: Exploring your dataset

If we rewind and go back to the learning objectives, you'll see that by the end of this deliverable, you should have formulated *4* research questions about your data that you may want to answer during your project. However, it may be handy to do some more exploration on your dataset of choice before creating these questions - by looking at the data, you may get more ideas. **Before you start this task, read all instructions carefully until you reach START HERE under Task 3**.

2.1 **(12 points)** Complete *4 out of the following 8 exercises* to dive deeper into your data. All datasets are different and therefore, not all of these tasks may make sense for your data - which is why you should only answer *4*.

Make sure that you're using dplyr and ggplot2 rather than base R for this task. Outside of this project, you may find that you prefer using base R functions for certain tasks, and that's just fine! But part of this project is for you to practice the tools we learned in class, which is dplyr and ggplot2.

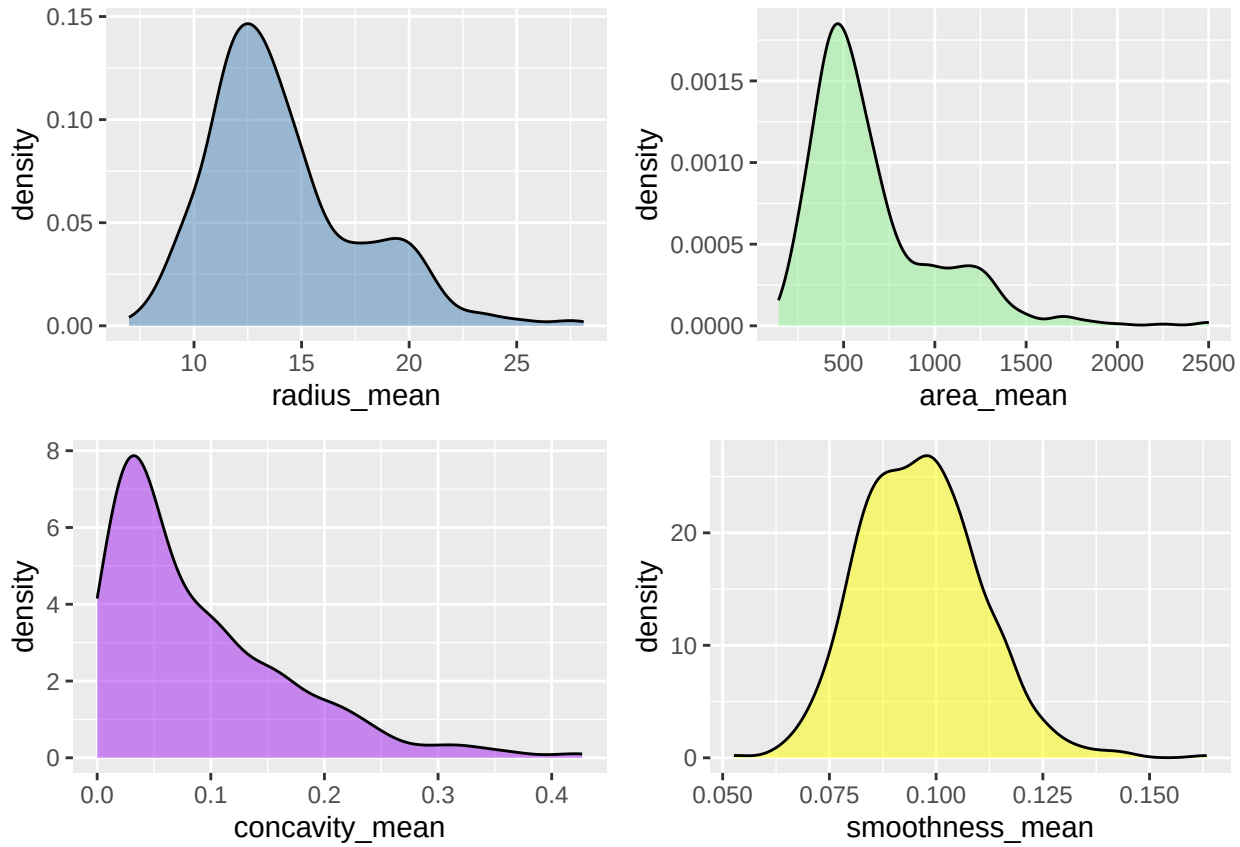
1. Plot the distribution of a numeric variable.
2. Create a new variable based on other variables in your data (only if it makes sense)
3. Investigate how many missing values there are per variable. Can you find a way to plot this?
4. Explore the relationship between 2 variables in a plot.
5. Filter observations in your data according to your own criteria. Think of what you'd like to explore - again, if this was the `titanic` dataset, I may want to narrow my search down to passengers born in a particular year...
6. Use a boxplot to look at the frequency of different observations within a single variable. You can do this for more than one variable if you wish!
7. Make a new tibble with a subset of your data, with variables and observations that you are interested in exploring.
8. Use a density plot to explore any of your variables (that are suitable for this type of plot).

Excercise 1: Distribution of numeric variables

```
# Use smoothed density esitimates to visualize the distributions of some key features
radius_mean_dist <- ggplot(cancer_sample, aes(x = radius_mean)) + geom_density(fill = "steelblue", alpha

area_mean_dist <- ggplot(cancer_sample, aes(x = area_mean)) + geom_density(fill = "lightgreen", alpha =

concavity_mean_dist <- ggplot(cancer_sample, aes(x = concavity_mean)) + geom_density(fill = "purple", al

smoothness_mean_dist <- ggplot(cancer_sample, aes(x = smoothness_mean)) + geom_density(fill = "yellow",

# install.packages("gridExtra")
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Arrange all plots in grids for better visualization
grid.arrange(radius_mean_dist, area_mean_dist,
             concavity_mean_dist, smoothness_mean_dist,
             ncol = 2, nrow = 2)
```



Excercise 2: Create a new variable based on other variables in your data

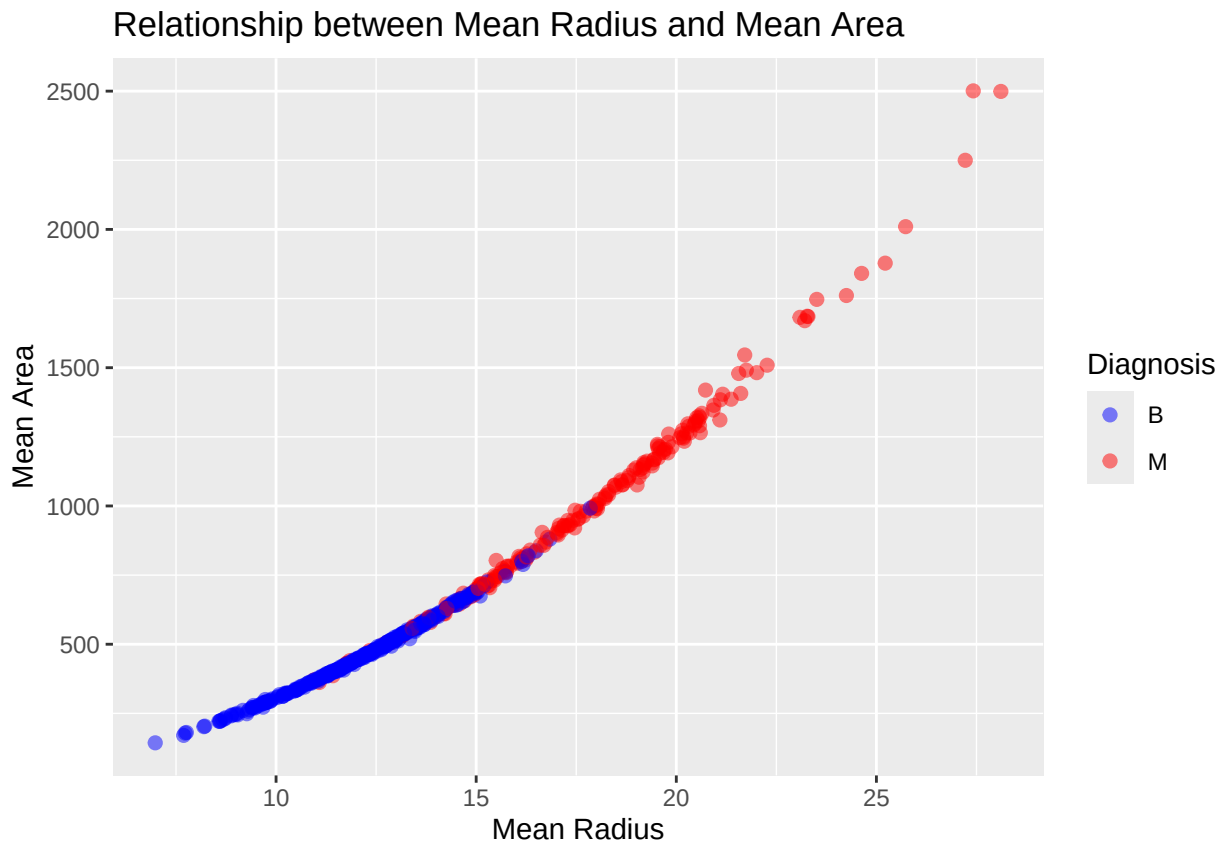
```
cancer_sample <- cancer_sample %>%
  mutate(
    # Create area-perimeter ratio to measure shape compactness
    # Higher values indicate more compact tumors
    area_perimeter_ratio = area_mean / perimeter_mean,
    # Classify tumors into size categories based on radius thresholds
    # Small: radius < 11, Medium: 11-15, Large: > 15
    size_category = case_when(
           radius_mean < 11 ~ "Small",
           radius_mean >= 11 & radius_mean < 15 ~ "Medium",
           radius_mean >= 15 ~ "Large"
         ))

cancer_sample %>%
  select(radius_mean, area_mean, perimeter_mean, area_perimeter_ratio, size_category) %>%
  head(10)
```

```
## # A tibble: 10 x 5
##    radius_mean area_mean perimeter_mean area_perimeter_ratio size_category
##          <dbl>     <dbl>          <dbl>                <dbl> <chr>
##  1        18.0     1001           123.                 8.15 Large
##  2        20.6     1326           133.                 9.98 Large
##  3        19.7     1203           130                  9.25 Large
##  4        11.4      386.           77.6                4.98 Medium
```

```
##  5        20.3     1297           135.                 9.60 Large
##  6        12.4      477.           82.6                5.78 Medium
##  7        18.2     1040           120.                 8.70 Large
##  8        13.7      578.           90.2                6.41 Medium
##  9        13        520.           87.5                5.94 Medium
## 10        12.5      476.           84.0                5.67 Medium
```

Excercise 4: Explore relationship between variables

```
# Explore relationship between radius_mean and area_mean, colored by diagnosis results
ggplot(cancer_sample, aes(x = radius_mean, y = area_mean, color = diagnosis)) +
  geom_point(alpha = 0.5, size = 2) +
  labs(title = "Relationship between Mean Radius and Mean Area",
       x = "Mean Radius",
       y = "Mean Area",
       color = "Diagnosis") +
  scale_color_manual(values = c("M" = "red", "B" = "blue"))
```



Excercise 7: Make a new tibble with a subset of your data

```
# Create a focused subset of malignant tumor data, extracting only key diagnostic measurements for mali

cancer_subset <- cancer_sample %>%
  select(ID, diagnosis,
         radius_mean, texture_mean, perimeter_mean, area_mean,
         concavity_mean, concave_points_mean) %>%
```

```
  filter(diagnosis == "M") %>%
  arrange(desc(radius_mean))


dim(cancer_subset)
```

```
## [1] 212   8
```

```
head(cancer_subset, 10)
```

```
## # A tibble: 10 x 8
##           ID diagnosis radius_mean texture_mean perimeter_mean area_mean
##        <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
##  1   8810703 M                28.1         18.5           188.      2499
##  2 911296202 M                27.4         26.3           187.      2501
##  3    873592 M                27.2         21.9           182.      2250
##  4    899987 M                25.7         17.5           174.      2010
##  5   8611555 M                25.2         24.9           172.      1878
##  6  91762702 M                24.6         21.6           166.      1841
##  7    865423 M                24.2         20.2           166.      1761
##  8     89812 M                23.5         24.3           155.      1747
##  9    878796 M                23.3         26.7           159.      1685
## 10   8712289 M                23.3         22.0           152.      1686
## # i 2 more variables: concavity_mean <dbl>, concave_points_mean <dbl>
```

2.2 **(4 points)** For each of the 4 exercises that you complete, provide a *brief explanation* of why you chose that exercise in relation to your data (in other words, why does it make sense to do that?), and sufficient comments for a reader to understand your reasoning and code.

Some explanations are attached in the comments within the code blocks above.

My exercise choices are 1,2,4, and 7.

Exercise 1: Examine the distribution of key tumor measurements (radius, area, concavity, smoothness) to identify patterns, detect outliers, and understand the range of values in the dataset.

Exercise 2: Create derived variables to capture tumor shape characteristics and size groupings, which may reveal non-linear relationships between measurements and diagnosis.

Exercise 4: Investigate the correlation between tumor radius and area, and examine how this relationship differs between malignant and benign diagnoses to identify diagnostic patterns.

Exercise 7: Generate a curated dataset focusing on core diagnostic measurements to facilitate targeted analysis of the most clinically significant tumor properties.

Task 3: Choose research questions

(4 points) So far, you have chosen a dataset and gotten familiar with it through exploring the data. You have also brainstormed one research question that interested you (Task 1.4). Now it's time to pick 4 research questions that you would like to explore in Milestone 2! Write the 4 questions and any additional comments below.

RQ1: Do malignant tumors have significantly larger measurements (radius, area, perimeter) compared to benign tumors?

RQ2: Are certain shape characteristics (concavity, symmetry, compactness) better indicators of malignancy than size measurements alone?

RQ3: What are the critical threshold values for radius and concavity that best separate malignant from benign tumors?

RQ4: Which combination of 2-3 tumor measurements provides the clearest separation between malignant and benign diagnoses?

Overall reproducibility/Cleanliness/Coherence Checklist

Coherence (0.5 points)

The document should read sensibly from top to bottom, with no major continuity errors. An example of a major continuity error is having a data set listed for Task 3 that is not part of one of the data sets listed in Task 1.

Error-free code (3 points)

For full marks, all code in the document should run without error. 1 point deduction if most code runs without error, and 2 points deduction if more than 50% of the code throws an error.

Main README (1 point)

There should be a file named `README.md` at the top level of your repository. Its contents should automatically appear when you visit the repository on GitHub.

Minimum contents of the README file:

- In a sentence or two, explains what this repository is, so that future-you or someone else stumbling on your repository can be oriented to the repository.
- In a sentence or two (or more??), briefly explains how to engage with the repository. You can assume the person reading knows the material from STAT 545A. Basically, if a visitor to your repository wants to explore your project, what should they know?

Once you get in the habit of making README files, and seeing more README files in other projects, you'll wonder how you ever got by without them! They are tremendously helpful.

Output (1 point)

All output is readable, recent and relevant:

- All Rmd files have been `knit`ted to their output md files.
- All knitted md files are viewable without errors on Github. Examples of errors: Missing plots, "Sorry about that, but we can't show files that are this big right now" messages, error messages from broken R code
- All of these output files are up-to-date – that is, they haven't fallen behind after the source (Rmd) files have been updated.
- There should be no relic output files. For example, if you were knitting an Rmd to html, but then changed the output to be only a markdown file, then the html file is a relic and should be deleted.

(0.5 point deduction if any of the above criteria are not met. 1 point deduction if most or all of the above criteria are not met.)

Our recommendation: right before submission, delete all output files, and re-knit each milestone's Rmd file, so that everything is up to date and relevant. Then, after your final commit and push to Github, CHECK on Github to make sure that everything looks the way you intended!

Tagged release (0.5 points)

You've tagged a release for Milestone 1.

Attribution

Thanks to Icíar Fernández Boyano for mostly putting this together, and Vincenzo Coia for launching.