# Topic ideas

**STA 210 - Project**

Team 2 - Alicia Gong, Ashley Chen, Abdel Shehata, Claire Tan

## Project idea 1

### Introduction and data

- State the source of the data set.
- Describe when and how it was originally collected (by the original data curator, not necessarily how you found the data)
- Describe the observations and the general characteristics being measured in the data

The Wine Quality dataset was collected from Vinho Verde wine Samples, from the North of Portugal. The data was originally donated in 2009 by Professor Cortez. The specific mechanism of the collection of the data was lab work done on different wines to measure their chemical attributites (like acidity etc.). The quality of the wine however was obtained through the average rating of three wine experts. The dataset is divided into two: Red wine and White wine. Red Wine has 1599 observations, and white wine has 4898 observations (each observation being a specific wine). Information about the wine include but are not limited to:PH,Density,Acidity, and alcohol content.

### Research question

- Describe a research question you're interested in answering using this data. Can Wine composition be used to predict it's quality?

  To answer this question, we would be looking at the input values from the physicochemical tests(acidity, Ph..etc) to determine if there is a relationship between those values and wine quality. If there is a relationship, then we would looking to model this relationship through linear regression, removing some of the corrlinear predictors.

  Another research question we could explore is: Can chemical composition of wine be used to predict it's color? (white or red)

To answer this question, we would be doing the same as above, however, with a focus on logistical regression. Thus, we would be looking at which qualities red and white wine share, and which they don't. Then using those to predict if a wine is red or white. Moreover, if there is a discrepancy in quality, then quality can be also used as a predictor.

## Glimpse of data

- Use the `glimpse` function to provide an overview of the dataset

```
library(tidyverse)
red_wine<-read.csv("data-1/winequality-red.csv", sep = ";")
white_wine<-read.csv("data-1/winequality-white.csv",sep = ";")
glimpse(red_wine)
```

```
Rows: 1,599
Columns: 12
$ fixed.acidity        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5~
$ volatile.acidity     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, ~
$ citric.acid          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0~
$ residual.sugar       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1,~
$ chlorides            <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069, ~
$ free.sulfur.dioxide  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, 16~
$ total.sulfur.dioxide <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 102,~
$ density              <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0~
$ pH                   <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3~
$ sulphates            <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47, 0~
$ alcohol              <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.~
$ quality              <int> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 7~
```

```
glimpse(white_wine)
```

```
Rows: 4,898
Columns: 12
$ fixed.acidity        <dbl> 7.0, 6.3, 8.1, 7.2, 7.2, 8.1, 6.2, 7.0, 6.3, 8.1,~
$ volatile.acidity     <dbl> 0.27, 0.30, 0.28, 0.23, 0.23, 0.28, 0.32, 0.27, 0~
$ citric.acid          <dbl> 0.36, 0.34, 0.40, 0.32, 0.32, 0.40, 0.16, 0.36, 0~
$ residual.sugar       <dbl> 20.70, 1.60, 6.90, 8.50, 8.50, 6.90, 7.00, 20.70,~
$ chlorides            <dbl> 0.045, 0.049, 0.050, 0.058, 0.058, 0.050, 0.045, ~
$ free.sulfur.dioxide  <dbl> 45, 14, 30, 47, 47, 30, 30, 45, 14, 28, 11, 17, 1~
$ total.sulfur.dioxide <dbl> 170, 132, 97, 186, 186, 97, 136, 170, 132, 129, 6~
```

```
$ density             <dbl> 1.0010, 0.9940, 0.9951, 0.9956, 0.9956, 0.9951, 0~
$ pH                  <dbl> 3.00, 3.30, 3.26, 3.19, 3.19, 3.26, 3.18, 3.00, 3~
$ sulphates           <dbl> 0.45, 0.49, 0.44, 0.40, 0.40, 0.44, 0.47, 0.45, 0~
$ alcohol             <dbl> 8.8, 9.5, 10.1, 9.9, 9.9, 10.1, 9.6, 8.8, 9.5, 11~
$ quality             <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 7, 5, 7, 6~
```

## Project idea 2

### Introduction and data

- State the source of the data set.
- Describe when and how it was originally collected (by the original data curator, not necessarily how you found the data)
- Describe the observations and the general characteristics being measured in the data

The dataset was originally posted by Noah Rippner on data.world.The dataset was collected through aggregation of a number of sources like census.gov or cancer.gov. The dataset has 3047 observations, with each observation representing a US county. Some respective information collected include but not limited to:median age,average household size, average deaths per year,birthrate, and cancer deathrate.

### Research question

- Describe a research question you're interested in answering using this data. A research question we want to explore is: Can a county's characteristics (racial composition, gender compositation. etc) be used to predict it's cancer deathrate? If so, What model is best for predicating?

  To answer this question, we would be looking at many factors from employment rate to racial composition of a county. Using those factor, and many others, we will be fitting models to determine which model is the parsimonious model.

### Glimpse of data

- Use the `glimpse` function to provide an overview of the dataset

```
library(readr)
cancer_reg <- read_csv("data-2/cancer_reg.csv")
glimpse(cancer_reg)
```

```
Rows: 3,047
Columns: 34
$ avgAnnCount            <dbl> 1397, 173, 102, 427, 57, 428, 250, 146, 88, 40~
$ avgDeathsPerYear       <dbl> 469, 70, 50, 202, 26, 152, 97, 71, 36, 1380, 3~
$ TARGET_deathRate       <dbl> 164.9, 161.3, 174.7, 194.8, 144.4, 176.0, 175.~
$ incidenceRate          <dbl> 489.8, 411.6, 349.7, 430.4, 350.1, 505.4, 461.~
```

```
$ medIncome               <dbl> 61898, 48127, 49348, 44243, 49955, 52313, 3778~
$ popEst2015              <dbl> 260131, 43269, 21026, 75882, 10321, 61023, 415~
$ povertyPercent          <dbl> 11.2, 18.6, 14.6, 17.1, 12.5, 15.6, 23.2, 17.8~
$ studyPerCap             <dbl> 499.74820, 23.11123, 47.56016, 342.63725, 0.00~
$ binnedInc               <chr> "(61494.5, 125635]", "(48021.6, 51046.4]", "(4~
$ MedianAge               <dbl> 39.3, 33.0, 45.0, 42.8, 48.3, 45.4, 42.6, 51.7~
$ MedianAgeMale           <dbl> 36.9, 32.2, 44.0, 42.2, 47.8, 43.5, 42.2, 50.8~
$ MedianAgeFemale         <dbl> 41.7, 33.7, 45.8, 43.4, 48.9, 48.0, 43.5, 52.5~
$ Geography               <chr> "Kitsap County, Washington", "Kittitas County,~
$ AvgHouseholdSize        <dbl> 2.5400, 2.3400, 2.6200, 2.5200, 2.3400, 2.5800~
$ PercentMarried          <dbl> 52.5, 44.5, 54.2, 52.7, 57.8, 50.4, 54.1, 52.7~
$ PctNoHS18_24            <dbl> 11.5, 6.1, 24.0, 20.2, 14.9, 29.9, 26.1, 27.3,~
$ PctHS18_24              <dbl> 39.5, 22.4, 36.6, 41.2, 43.0, 35.1, 41.4, 33.9~
$ PctSomeCol18_24         <dbl> 42.1, 64.0, NA, 36.1, 40.0, NA, NA, 36.5, NA, ~
$ PctBachDeg18_24         <dbl> 6.9, 7.5, 9.5, 2.5, 2.0, 4.5, 5.8, 2.2, 1.4, 7~
$ PctHS25_Over            <dbl> 23.2, 26.0, 29.0, 31.6, 33.4, 30.4, 29.8, 31.6~
$ PctBachDeg25_Over       <dbl> 19.6, 22.7, 16.0, 9.3, 15.0, 11.9, 11.9, 11.3,~
$ PctEmployed16_Over      <dbl> 51.9, 55.9, 45.9, 48.3, 48.2, 44.1, 51.8, 40.9~
$ PctUnemployed16_Over    <dbl> 8.0, 7.8, 7.0, 12.1, 4.8, 12.9, 8.9, 8.9, 10.3~
$ PctPrivateCoverage      <dbl> 75.1, 70.2, 63.7, 58.4, 61.6, 60.0, 49.5, 55.8~
$ PctPrivateCoverageAlone <dbl> NA, 53.8, 43.5, 40.3, 43.9, 38.8, 35.0, 33.1, ~
$ PctEmpPrivCoverage      <dbl> 41.6, 43.6, 34.9, 35.0, 35.1, 32.6, 28.3, 25.9~
$ PctPublicCoverage       <dbl> 32.9, 31.1, 42.1, 45.3, 44.0, 43.2, 46.4, 50.9~
$ PctPublicCoverageAlone  <dbl> 14.0, 15.3, 21.1, 25.0, 22.7, 20.2, 28.7, 24.1~
$ PctWhite                <dbl> 81.78053, 89.22851, 90.92219, 91.74469, 94.104~
$ PctBlack                <dbl> 2.5947283, 0.9691025, 0.7396734, 0.7826260, 0.~
$ PctAsian                <dbl> 4.82185710, 2.24623259, 0.46589818, 1.16135867~
$ PctOtherRace            <dbl> 1.84347853, 3.74135153, 2.74735831, 1.36264318~
$ PctMarriedHouseholds    <dbl> 52.85608, 45.37250, 54.44487, 51.02151, 54.027~
$ BirthRate               <dbl> 6.1188310, 4.3330956, 3.7294878, 4.6038408, 6.~
```

## Project idea 3

### Introduction and data

The data was originally uploaded by the user Dgomonov on kaggle.com and sourced from the Airbnb website. The dataset includes listing activities and metrics for New York City, New York in 2019. The data file includes information on hosts, geographical availability, number of reviews, price, neighborhood, availability, and more. There are 48895 observations (each airbnb host) and 16 variables being measured in the data.

### Research question

- Describe a research question you're interested in answering using this data.
  - Which features or predictions (locations, reviews, cost, etc.) have the greatest influence on airbnb reservations?
  - To answer this question, we would use multiple linear regression and a variety of methods to determine which factors or variables are integral in increasing the likelihood that a customer books an Airbnb and fit the "best" model.

```
air_bnb <- read_csv("data-3/AB_NYC_2019.csv")
glimpse(air_bnb)
```

```
Rows: 48,895
Columns: 16
$ id                                <dbl> 2539, 2595, 3647, 3831, 5022, 5099, 512~
$ name                              <chr> "Clean & quiet apt home by the park", "~
$ host_id                           <dbl> 2787, 2845, 4632, 4869, 7192, 7322, 735~
$ host_name                         <chr> "John", "Jennifer", "Elisabeth", "LisaR~
$ neighbourhood_group               <chr> "Brooklyn", "Manhattan", "Manhattan", "~
$ neighbourhood                     <chr> "Kensington", "Midtown", "Harlem", "Cli~
$ latitude                          <dbl> 40.64749, 40.75362, 40.80902, 40.68514,~
$ longitude                         <dbl> -73.97237, -73.98377, -73.94190, -73.95~
$ room_type                         <chr> "Private room", "Entire home/apt", "Pri~
$ price                             <dbl> 149, 225, 150, 89, 80, 200, 60, 79, 79,~
$ minimum_nights                    <dbl> 1, 1, 3, 1, 10, 3, 45, 2, 2, 1, 5, 2, 4~
$ number_of_reviews                 <dbl> 9, 45, 0, 270, 9, 74, 49, 430, 118, 160~
$ last_review                       <date> 2018-10-19, 2019-05-21, NA, 2019-07-05~
$ reviews_per_month                 <dbl> 0.21, 0.38, NA, 4.64, 0.10, 0.59, 0.40,~
$ calculated_host_listings_count    <dbl> 6, 2, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 3, ~
$ availability_365                  <dbl> 365, 355, 365, 194, 0, 129, 0, 220, 0, ~
```