EDA

Patrick Seebold

```
Warning: package 'tidymodels' was built under R version 4.3.3

-- Attaching packages ------------------------------------- tidymodels 1.2.0 --

v broom        1.0.5     v recipes      1.1.0
v dials        1.3.0     v rsample      1.2.1
v dplyr        1.1.4     v tibble       3.2.1
v ggplot2      3.5.1     v tidyr        1.3.1
v infer        1.0.7     v tune         1.2.1
v modeldata    1.4.0     v workflows    1.1.4
v parsnip      1.2.1     v workflowsets 1.1.0
v purrr        1.0.2     v yardstick    1.3.1

Warning: package 'dials' was built under R version 4.3.3

Warning: package 'scales' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'infer' was built under R version 4.3.3

Warning: package 'modeldata' was built under R version 4.3.3

Warning: package 'parsnip' was built under R version 4.3.3

Warning: package 'purrr' was built under R version 4.3.3
```

```
Warning: package 'recipes' was built under R version 4.3.3

Warning: package 'rsample' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'tune' was built under R version 4.3.3

Warning: package 'workflows' was built under R version 4.3.3

Warning: package 'workflowsets' was built under R version 4.3.3

Warning: package 'yardstick' was built under R version 4.3.3

-- Conflicts --------------------------------------- tidymodels_conflicts() --
x purrr::discard() masks scales::discard()
x dplyr::filter()  masks stats::filter()
x dplyr::lag()     masks stats::lag()
x recipes::step()  masks stats::step()
* Search for functions across packages at https://www.tidymodels.org/find/
```

In this project, we will construct a predictive model based on diabetes status. This means we will create a model that, given some information about a new patient, we will be able to predict whether they will have diabetes or not. We'll use a publicly available data set for this (diabetes_binary_health_indicators_BRFSS2015.csv), and will then deploy the model in a docker file to allow for easy sharing/calling of the program. This document will handle the EDA, and the second file will handle the training/testing of the model. First, let's grab the data:

head(data)

sum(is.na(data))

```
data = read.csv("diabetes_binary_health_indicators_BRFSS2015.csv")
```

Next, let's take a look at the head of the file to confirm everything loaded in. Then we'll check for missing data, and finally adjust the type of our variables of interest. Rather than use the standard predictive variables like blood pressure, I've decided to see whether a person's self-perceived health may be indicative of a health issue, so our predicting variables will be Education, Sex, GenHlth, MentHlth, PhysHlth.

Note that GenHlth is a 5 point scale of participant's self-reported health, 1 being excellent, 5 being poor. MenHlth and PhysHlth are how many days of past 30 days an individual has struggled with Mental or Physical health respectively. We'll treat GenHlth as a factor and MenHlth/PhysHlth as numeric.

```
head(data) # looks good
```

```
  Diabetes_binary HighBP HighChol CholCheck BMI Smoker Stroke
1               0      1        1         1  40      1      0
2               0      0        0         0  25      1      0
3               0      1        1         1  28      0      0
4               0      1        0         1  27      0      0
5               0      1        1         1  24      0      0
6               0      1        1         1  25      1      0
  HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
1                    0            0      0       1                 0
2                    0            1      0       0                 0
3                    0            0      1       0                 0
4                    0            1      1       1                 0
5                    0            1      1       1                 0
6                    0            1      1       1                 0
  AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
1             1           0       5       18       15        1   0   9
2             0           1       3        0        0        0   0   7
3             1           1       5       30       30        1   0   9
4             1           0       2        0        0        0   0  11
5             1           0       2        3        0        0   0  11
6             1           0       2        0        2        0   1  10
  Education Income
1         4      3
2         6      1
3         4      8
4         3      6
5         5      4
6         6      8
```

```
sum(is.na(data)) # no missing data, sweet!
```

```
[1] 0
```

```r
# before we continue, let's subset our data to just the variables we plan to explore:
data_sub = data[, c("Diabetes_binary", "Education", "Sex", "GenHlth", "MentHlth", "PhysHlth")]
summary(data_sub)
```

```
 Diabetes_binary    Education          Sex            GenHlth     
 Min.   :0.0000   Min.   :1.00   Min.   :0.0000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:4.00   1st Qu.:0.0000   1st Qu.:2.000  
 Median :0.0000   Median :5.00   Median :0.0000   Median :2.000  
 Mean   :0.1393   Mean   :5.05   Mean   :0.4403   Mean   :2.511  
 3rd Qu.:0.0000   3rd Qu.:6.00   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :1.0000   Max.   :6.00   Max.   :1.0000   Max.   :5.000  
    MentHlth         PhysHlth      
 Min.   : 0.000   Min.   : 0.000  
 1st Qu.: 0.000   1st Qu.: 0.000  
 Median : 0.000   Median : 0.000  
 Mean   : 3.185   Mean   : 4.242  
 3rd Qu.: 2.000   3rd Qu.: 3.000  
 Max.   :30.000   Max.   :30.000  
```

```r
typeof(data_sub$MentHlth)
```

```
[1] "double"
```

```r
typeof(data_sub$PhysHlth)
```

```
[1] "double"
```

```r
data_sub$Diabetes_binary = factor(data_sub$Diabetes_binary, levels = c('1','0'),
                                  labels = c("Diabetes", "No Diabetes"))
data_sub$Sex = factor(data_sub$Sex, levels = c('0','1'),
                      labels = c("Female","Male"))
data_sub$Education = factor(data_sub$Education,
                            levels = c('1','2','3','4','5','6'),
                            labels = c("Never attended school or only kindergarten",
                            "Grades 1 through 8 (Elementary)",
                            "Grades 9 through 11 (Some high school)",
                            "Grade 12 or GED (High school graduate)",
                            "College 1 year to 3 years (Some college or technical school)"
                            "College 4 years or more (College graduate)"))
```

```
data_sub$GenHlth = factor(data_sub$GenHlth, levels = c('1','2','3','4','5'),
                          labels = c("Excellent", "Very Good", "Good",
                                     "Fair","Poor"))
levels(data_sub$Diabetes_binary)
```

```
[1] "Diabetes"    "No Diabetes"
```

```
levels(data_sub$Sex)
```

```
[1] "Female" "Male"
```

```
levels(data_sub$Education)
```

```
[1] "Never attended school or only kindergarten"
[2] "Grades 1 through 8 (Elementary)"
[3] "Grades 9 through 11 (Some high school)"
[4] "Grade 12 or GED (High school graduate)"
[5] "College 1 year to 3 years (Some college or technical school)"
[6] "College 4 years or more (College graduate)"
```

```
levels(data_sub$GenHlth)
```

```
[1] "Excellent" "Very Good" "Good"      "Fair"      "Poor"
```

Great, now we have our factors set and we've confirmed that our numeric variables are the proper type. We can now do some numeric summaries to get a look at how our variables relate to each other:

```
data_sub |> # let's see how Mental and Physical health vary across Male and Female
  group_by(Sex) |>
  summarize(Mmean = mean(MentHlth), Msd = sd(MentHlth), Pmean = mean(PhysHlth), Psd = sd(Phys
```

```
# A tibble: 2 x 5
  Sex    Mmean   Msd Pmean   Psd
  <fct>  <dbl> <dbl> <dbl> <dbl>
1 Female  3.72  7.87  4.58  8.90
2 Male    2.51  6.72  3.82  8.46
```

```
data_sub |> # Same summary but across education groups
  group_by(Education) |>
  summarize(Mmean = mean(MentHlth), Msd = sd(MentHlth), Pmean = mean(PhysHlth), Psd = sd(Phys
```

```
# A tibble: 6 x 5
  Education                                              Mmean   Msd Pmean   Psd
  <fct>                                                  <dbl> <dbl> <dbl> <dbl>
1 Never attended school or only kindergarten              6.58 10.8   8.85 12.4
2 Grades 1 through 8 (Elementary)                         5.16  9.72  8.35 11.6
3 Grades 9 through 11 (Some high school)                  5.32  9.79  7.92 11.3
4 Grade 12 or GED (High school graduate)                  3.67  8.10  5.30  9.72
5 College 1 year to 3 years (Some college or technical ~  3.65  7.93  4.68  9.13
6 College 4 years or more (College graduate)              2.33  6.07  2.85  6.99
```

```
data_sub |> # Finally, summarizing these health variables grouping by diabetes
  group_by(Diabetes_binary) |>
  summarize(Mmean = mean(MentHlth), Msd = sd(MentHlth), Pmean = mean(PhysHlth), Psd = sd(Phys
```

```
# A tibble: 2 x 5
  Diabetes_binary Mmean   Msd Pmean   Psd
  <fct>           <dbl> <dbl> <dbl> <dbl>
1 Diabetes         4.46  8.95  7.95 11.3
2 No Diabetes      2.98  7.11  3.64  8.06
```

```
# There is some correlation between physical and mental health variables
cor(data_sub$MentHlth, data_sub$PhysHlth)
```

```
[1] 0.3536189
```

Our summaries suggest some interesting trends! First, it looks like females may be more likely to have Mental and Physical health days than Males, although these are relatively small differences. More education is similarly associated with fewer mental and physical health days, although it is entirely possible that health problems may account for why some participants did not attain higher education. We also see that the diabetes group reports poorer health outcomes for both mental and physical health, with physical health showing an average of more ~4.5 days of physical health problems in the past 30 days vs the no diabetes group. We also see a fair correlation between the mental and physical health variables.

Since we have only a few numeric variables, we can't meaningfully takes means and standard deviations for all our variables of interest. For our factor variables we will use contingency tables to get a better idea of how these things are working:

```
# Mental and Physical health tend to score higher when General Health is reported as better
table(data_sub$GenHlth,data_sub$PhysHlth)
```

	0	1	2	3	4	5	6	7	8	9	10
Excellent	38274	1872	1690	850	387	545	77	381	50	15	252
Very Good	65983	5177	5714	2743	1312	2041	293	1303	189	33	978
Good	45984	3554	5595	3425	1835	3121	519	1707	292	75	2092
Fair	9130	741	1638	1301	881	1700	369	981	231	43	1923
Poor	681	44	127	176	127	215	72	166	47	13	350

	11	12	13	14	15	16	17	18	19	20	21
Excellent	7	22	3	179	98	4	2	7	1	61	49
Very Good	14	95	20	643	507	16	15	14	2	295	135
Good	11	183	14	956	1481	23	17	37	6	751	222
Fair	19	217	27	656	2144	47	37	72	10	1453	196
Poor	9	61	4	153	686	22	25	22	3	713	61

	22	23	24	25	26	27	28	29	30
Excellent	3	2	2	23	1	6	15	5	416
Very Good	7	6	11	97	9	6	40	19	1367
Good	13	12	14	244	12	19	82	32	3318
Fair	26	25	26	567	22	34	194	68	6792
Poor	21	11	19	405	25	34	191	91	7507

```
table(data_sub$GenHlth,data_sub$MentHlth)
```

	0	1	2	3	4	5	6	7	8	9	10
Excellent	36738	1529	1996	1003	487	1087	98	359	61	7	533
Very Good	65610	3679	5101	2732	1371	2957	317	977	202	21	1692
Good	51776	2480	4073	2392	1212	2986	328	1068	183	29	2087
Fair	16897	702	1496	982	541	1527	172	524	151	26	1469
Poor	4659	148	388	272	178	473	73	172	42	8	592

	11	12	13	14	15	16	17	18	19	20	21
Excellent	1	34	1	97	346	9	4	5	2	172	22
Very Good	11	98	11	309	1225	15	7	17	3	628	37
Good	17	141	20	397	1639	28	19	28	4	1037	68
Fair	11	93	5	270	1510	25	17	35	5	992	70
Poor	1	32	4	94	785	11	7	12	2	535	30

```
              22    23    24    25    26    27    28    29    30
  Excellent    6     2     1    55     2     5    22    18   597
  Very Good   13     9     9   187    10    20    57    41  1718
  Good        13    14    12   322    14    23   104    39  3093
  Fair        18    11     8   379    12    15    72    35  3500
  Poor        13     2     3   245     7    16    72    25  3180
```

```
# Higher proportion of cases of diabetes at poorer levels of general health
table(data_sub$Diabetes_binary, data_sub$GenHlth)
```

```
             Excellent Very Good  Good  Fair  Poor
  Diabetes        1140      6381 13457  9790  4578
  No Diabetes    44159     82703 62189 21780  7503
```

```
# Difficult to make any firm conclusion from education/sex/diabetes table, we'll do some gra
table(data_sub$Diabetes_binary, data_sub$Sex, data_sub$Education)
```

```
, ,  = Never attended school or only kindergarten


             Female  Male
  Diabetes       30    17
  No Diabetes    72    55

, ,  = Grades 1 through 8 (Elementary)


             Female  Male
  Diabetes      677   506
  No Diabetes  1504  1356

, ,  = Grades 9 through 11 (Some high school)


             Female  Male
  Diabetes     1377   919
  No Diabetes  4135  3047

, ,  = Grade 12 or GED (High school graduate)
```

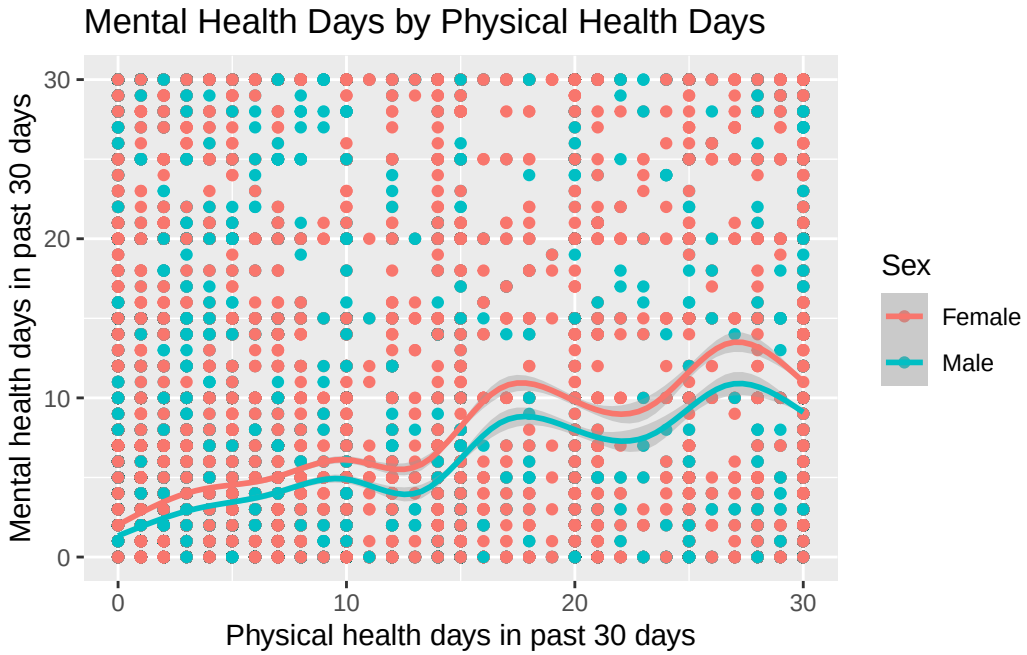
```
             Female  Male
  Diabetes      6106  4960
  No Diabetes  29014 22670

, ,  = College 1 year to 3 years (Some college or technical school)


             Female  Male
  Diabetes      5683  4671
  No Diabetes  35539 24017

, ,  = College 4 years or more (College graduate)


             Female  Male
  Diabetes      4538  5862
  No Diabetes  53299 43626
```

Overall, it does look like self-reported general health may differ across diabetes vs non-diabetes patients. Mental and Physical health variables also appear to have a relationship with general health rating, which is what we'd expect to see here. It's also interesting to note that the number of poor health days tend to decrease as we move from Day 1 to Day 29, but there is an unexpectedly high number of people reporting 30 Days of issues. This suggests that there is a subset of participants that experience some sort of daily chronic issue.

We'll next do some plots so we can visualize some of these relationships:

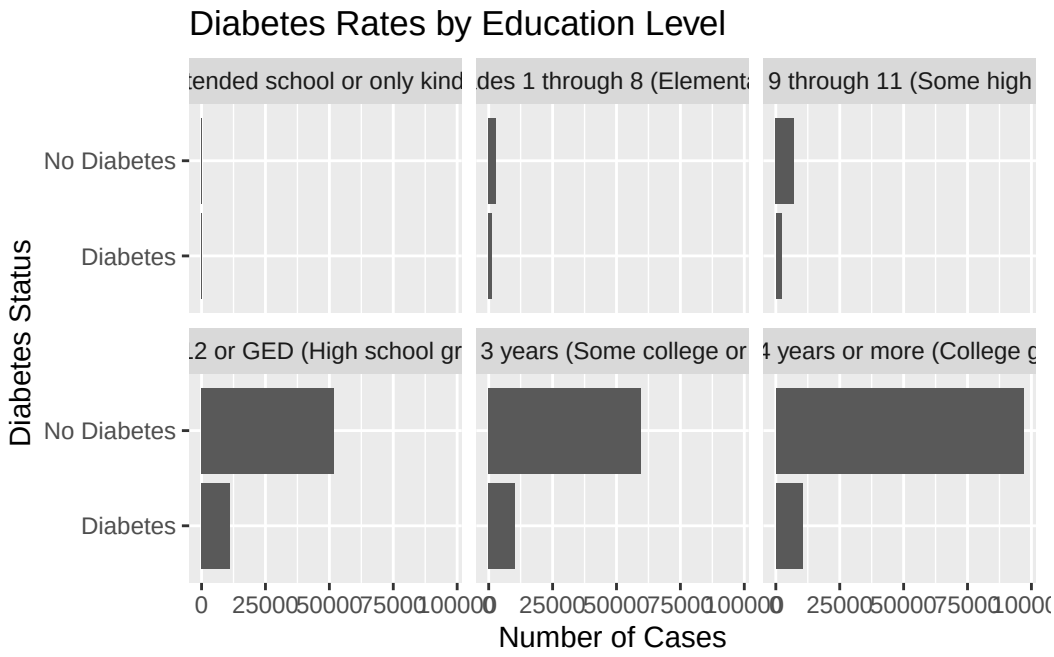
```
# scatter plots of Physical*Mental Health across sex
g = ggplot(data_sub, aes(y = MentHlth, x = PhysHlth, color = Sex))
g + geom_point() + geom_smooth() +
labs(title = "Mental Health Days by Physical Health Days", x = "Physical health days in past
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



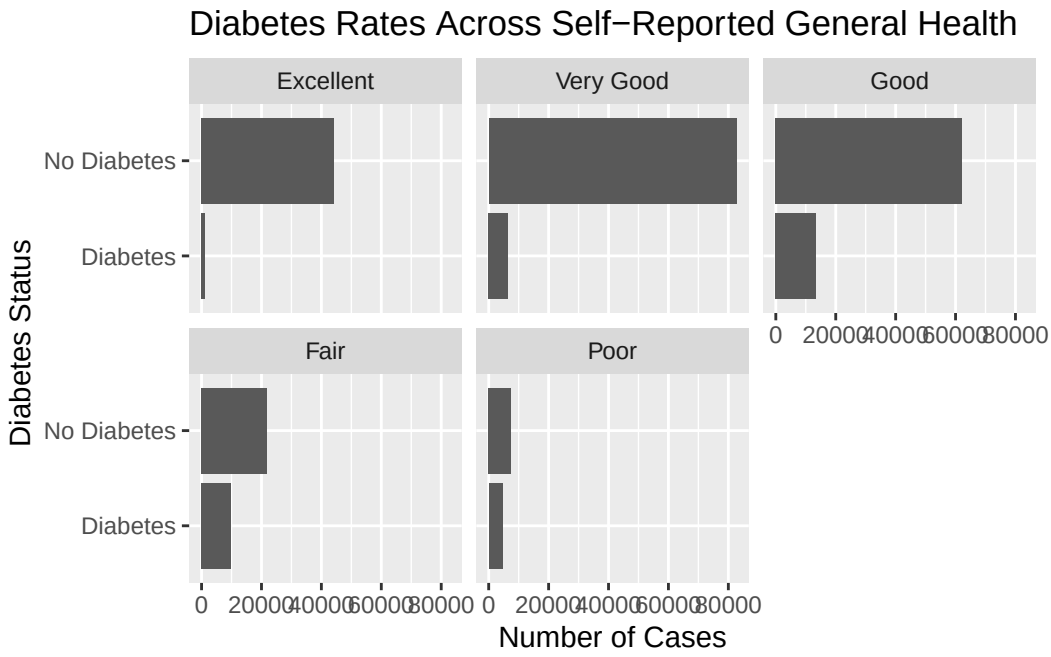
We can visualize the positive correlation between the Mental and Physical Health, as expected. Next, let's visualize education and diabetes since our table was tricky to interpret:

```
# Bar plots of general health and diabetes by education level
g = ggplot(data_sub, aes(y = Diabetes_binary))
g + geom_bar() +
labs(title = "Diabetes Rates by Education Level", x = "Number of Cases", y = "Diabetes Status
facet_wrap(~Education)
```



It does look like we might have a lower proportion of diabetes in higher education groups, since the number of diabetes cases appears similar between the highest three education groups while the total sample size of each education group gets progressively larger. Finally, let's see how General Health plots with diabetes:

```
# Bar plots of diabetes rate by general health rating
g = ggplot(data_sub, aes(y = Diabetes_binary))
g + geom_bar() +
labs(title = "Diabetes Rates Across Self-Reported General Health", x = "Number of Cases", y =
facet_wrap(~GenHlth)
```



Here, we see that there is a higher proportion of diabetes in groups that report lower general health. This matches with the intuition that individuals with diabetes may consider themselves to be less healthy on average. Now that we have gotten an idea for how the variables relate, we can move onto training our models in our second document.

Click here for the Modeling Page