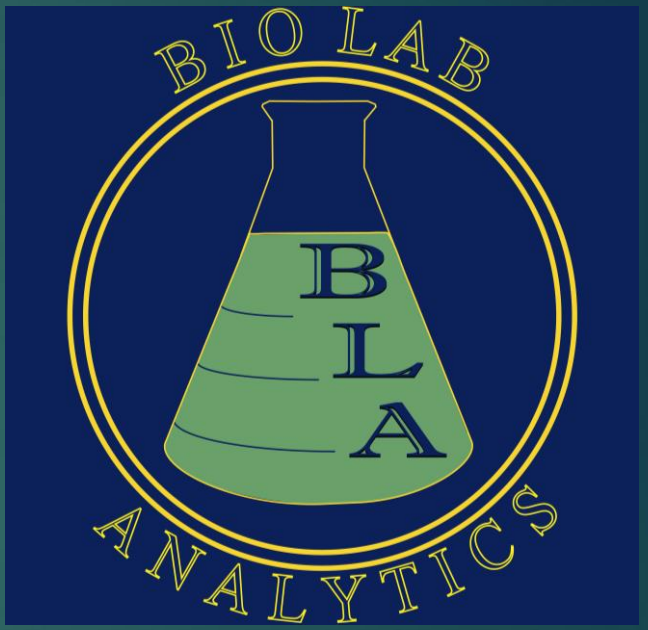
Data Science for Biologists
Regression & Classification

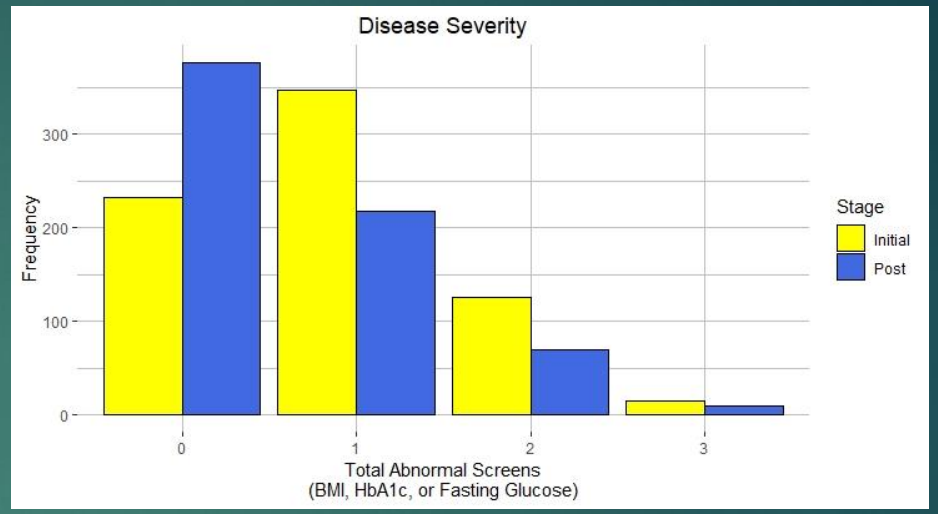


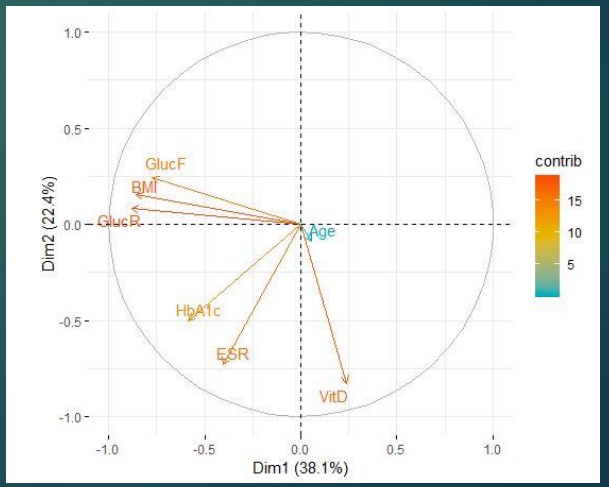
PAUL SOCHACKI, MS

06/10/2019

DS Basics- Overview

- Regular Expressions
- Regression Techniques
  - Linear
  - Logistic
  - Poisson
- Grouping/Classification Techniques
  - Principle Components Analysis
  - Correspondence Analysis





DS Basics- Regular Expressions

- Regular expressions (or RegEx) can be used to quickly manipulate & search string data.
- Can specify exact match, wildcards, and other parameters for search
- Nearly universal – most analytics platforms support some RegEx

- Reserved Characters:

{} [] () ^ $. | * + ? \ - #

	PTID	DNA_ID
1	C099561457	CCGGTCGAGTTCTAACGCCTAGTCCAAATCCGCTAGGCAT...
2	C099249488	CCGGTCGGCTGGAATCGGGAATTGGAGTCCCGATTCTAGA...
3	C099216418	CCGGTCGGGTTCTACGGAATTGAGCTGTGATAATTTTTCGT...
4	C099165692	CCGGTCGTATCCAGAGCGGCGACTCAATCCACCATCGATC...
5	C098473923	CCGGTCGCTTAGTACCTACACTCGACCCTAGAAGAACGCG...
6	C098341356	CCGGTCGAGAAGCGAATGAACTATCCTGCGTATCAAACAA...
7	C097923467	CCGGTCGAGCGTCGCGTAAAGCACTTCATTCTAGACGCCC...
8	C097813333	CCGGTCGACTCTCATCTCAACCGGGATTCTCACGTTGAAGA...
9	C097234826	CCGGTCGCAGACATCCTCTGCGCGTGTGTGACTGCTTTGT...
10	C097169857	CCGGTCGTGCGACACGTTTATTTTCCCATTGGCAAAGCAA...
11	C097154747	CCGGTCGTTGAGCAGGGCAGGCTACGAGACGTAGCGGT...
12	C096742931	CCGGTCGCTAAGAGGCACTTCTTTGGGATACGTAATATGTT...

DS Basics- Regular Expressions

- Several different functions for this in R.
- Great for web addresses, genetic sequences

```
> grep(pattern = "GGGTTT{2}", x=Dems_DNA$DNA_ID, value = TRUE)
 ## value option returns string
[1] "CCGGTCGACTAAAAACCGTGAGGGTTTTAAGGATGCCCGAGGACTCGAGGCTGGCC"
[2] "CCGGTCGCACCTCCTACTCTCGCATTCAGGGTTTTCGTTACTCATGAACGCTGGCC"
```

- grep() – returns index value for first exact match
- grepl() – returns logic value for entire vector
- regrexpr() – returns character position for first match
- gregexpr() – returns character position for multiple matches

```
[685] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[694] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
[703] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[712] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[721] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[730] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[739] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[748] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[757] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
[766] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

DS Basics- Linear Regression

Use Case:

Given a predictor or independent variable (X), predict a response or dependent variable (Y). Predictions are based on historical data.

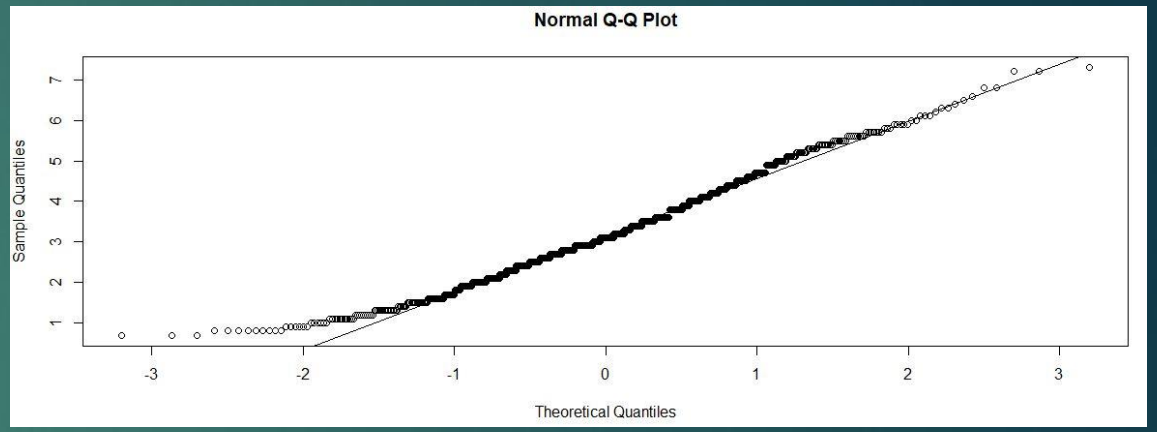
Key Assumptions:

- Quantitative, continuous predictor and response variables
- *Normally distributed* residuals among the response variable
- No outliers in the dataset (?)

Tests for normal distribution:

- Q-Q Plot – markers should line up in a diagonal, LL to UR
- Wilkes-Shapiro test – W statistic >= 0.80 marginal, >=0.90 good

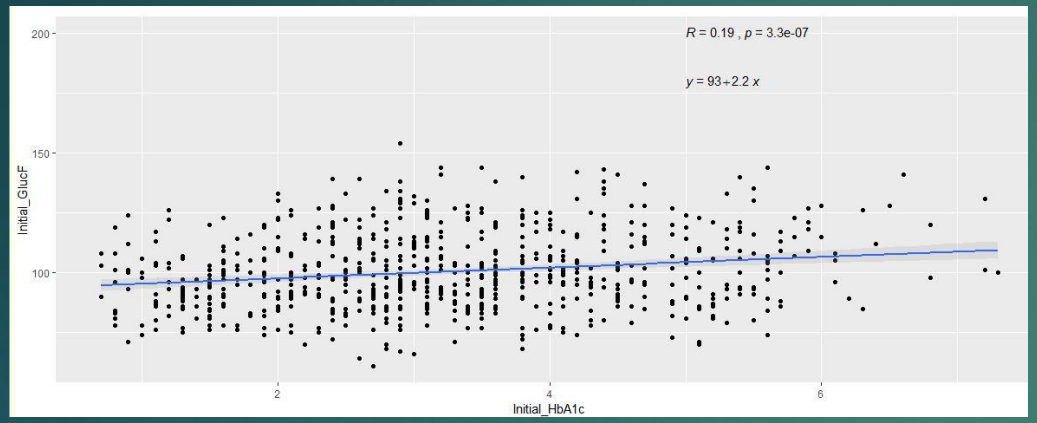
```
> qqnorm(ptLabs$Initial_HbA1c)
> qqline(ptLabs$Initial_HbA1c)
```

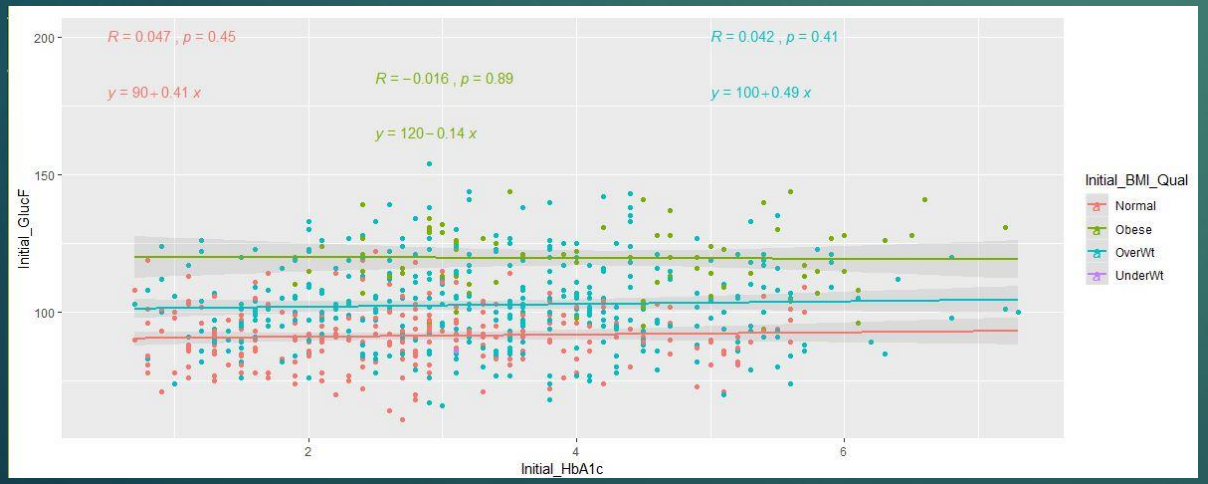



```
> shapiro.test(ptLabs$Initial_HbA1c)

        Shapiro-Wilk normality test

data:  ptLabs$Initial_HbA1c
W = 0.98169, p-value = 7.913e-08
```

DS Basics- Linear Regression Example

- **Question:** Is there a relationship Initial Hemoglobin A1c level and Fasting Glucose?





- Results often visualized as scatter with a fitted line.
- R is the percent of variance explained for the model; the bigger the better.
- p is the statistical significance of the model; the smaller the better.
- Confidence intervals often "flared" in areas of regression curve with low representation.

DS Basics- Logistic Regression

- **Use Case:**

  Given a predictor or independent variable (X), predict a response or dependent variable (Y). Predictions are based on historical data.

Key Assumptions:

  - Binomial (yes/no) response variable
  - Continuous, binomial, or categorial predictor
  - Adequate cell-size for predictive modeling

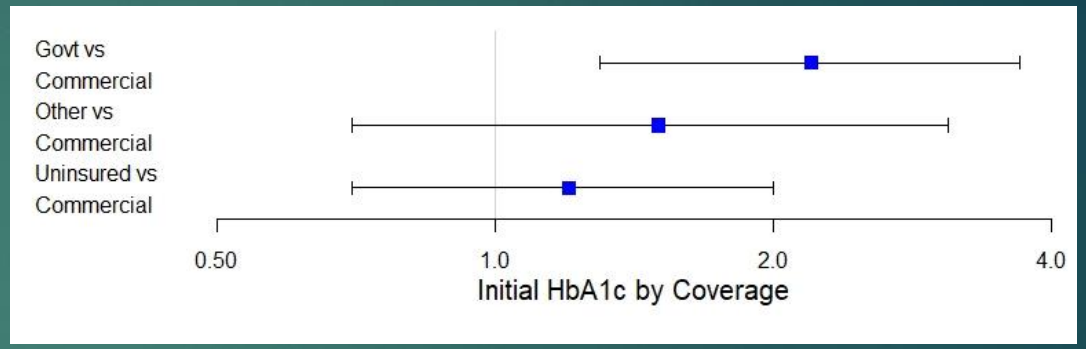
```
> table(ptLabs2$Initial_HbA1c_Qual, ptLabs2$Coverage, useNA="ifany")

      COMM GOVT OTHER UNINS
  Diab    6    6     1     1
  Norm  335   77    42   139
  Risk   51   23    10    28
```

- **Cell-size issues:**
  - Generally, absolute minimal cell-size N=5
  - *Workaround #1*: re-binning the predictor, if categorical variable.
  - *Workaround #2*: adjusting threshold for response, consistent with client technical specs.

DS Basics- Logistic Regression Example

- **Question:** Is insurance coverage associated with HbA1c, a diagnostic test for diabetes?

```
> model2 <- glm(Initial_HbA1c_Qual~Coverage, data=ptLabs2, family = "binomial")
> summary(model2)

Call:
glm(formula = Initial_HbA1c_Qual ~ Coverage, family = "binomial",
    data = ptLabs2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7995  -0.6156  -0.5606  -0.5606   1.9638

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.7711     0.1433 -12.361  < 2e-16 ***
CoverageGOVT   0.7946     0.2608   3.047  0.00231 **
CoverageOTHER  0.4313     0.3678   1.173  0.24088
CoverageUNINS  0.2039     0.2494   0.818  0.41363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 667.38  on 718  degrees of freedom
Residual deviance: 658.19  on 715  degrees of freedom
AIC: 666.19
```




```
> expb <- exp(coef(model2))
> print(expb)
  (Intercept)  CoverageGOVT CoverageOTHER CoverageUNINS
    0.1701493     2.2134883     1.5392648     1.2261770
> intexp <- exp(confint(model2))
Waiting for profiling to be done...
> print(intexp)
                   2.5 %    97.5 %
(Intercept)    0.1272264 0.2233389
CoverageGOVT   1.3165900 3.6705050
CoverageOTHER  0.7178786 3.0735936
CoverageUNINS  0.7443469 1.9850073
```

- Results can be visualized as a Forest plot.
- The markers are called “point estimates”, predictions of risk.
- The bars are confidence intervals, or the margin of error surrounding the estimates.

DS Basics- Poisson Regression

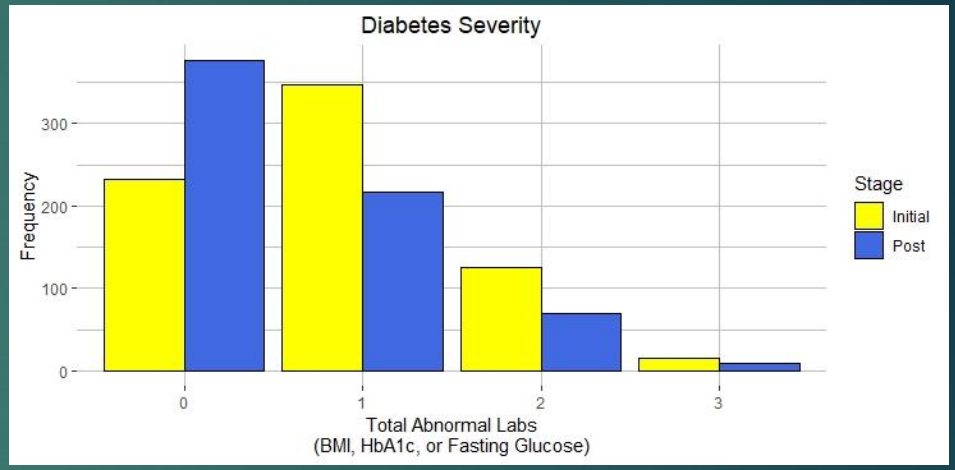
- **Use Case:**

  Predict whether counts of a binomial outcome (Y) are associated with a categorical variable.

Key Assumptions:

  - Counts of binomial (yes/no) response variable
  - Categorical predictor variable
  - Adequate cell-size for predictive modeling

- **Cell-size issues:**
  - *Workaround #1*: re-binning the counts, if applicable.

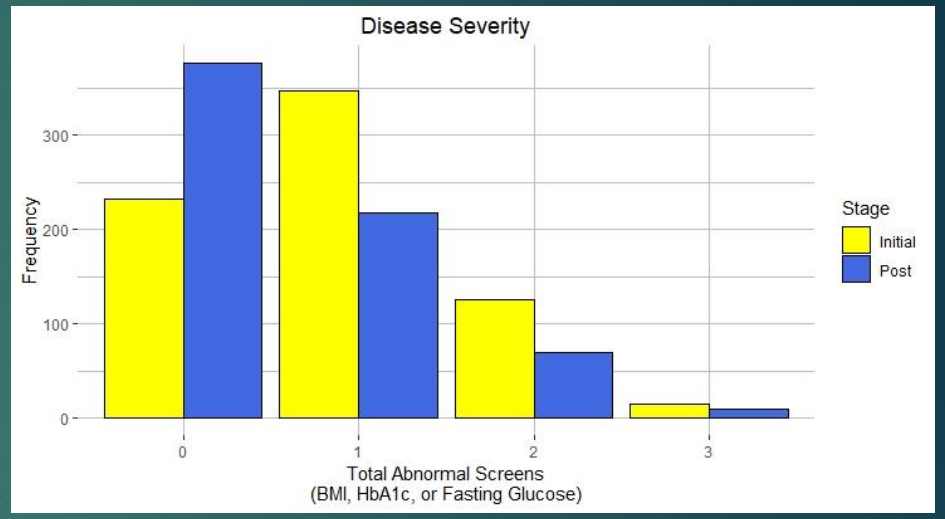


DS Basics- Poisson Regression Example

- **Question:** Did error counts significantly decrease between Initial and Post treatment stages?

```
> PR_model <- glm(Severity~Time, family="poisson",
+                 data= ptLabsPRl)
> summary(PR_model)

Call:
glm(formula = Severity ~ Time, family = "poisson", data = ptLabsPRl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3363  -1.0671   0.1112   0.5151   2.2606

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.11327    0.03947  -2.870   0.0041 **
TimePost_DiabSev -0.45008    0.06462  -6.965 3.28e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1321.9  on 1389  degrees of freedom
Residual deviance: 1271.9  on 1388  degrees of freedom
AIC: 2982.8
```

```
> exp(PR_model$coefficients) ## display estimate
     (Intercept) TimePost_DiabSev
       0.8929068        0.6375800
> exp(confint(PR_model)) ## display CI for the estimate
Waiting for profiling to be done...
                      2.5 %    97.5 %
(Intercept)       0.8256053 0.9637672
TimePost_DiabSev 0.5613295 0.7232039
```



The "Post" estimate is -0.45, which translates to an incidence rate of 0.64 (exp(-0.45)=0.64). This means there were 36% fewer counts with the intervention.

Principal Components Analysis (PCA)

- **Use Case:**
  Identify commonalities among quantitative variables.

Key Assumptions:

  - Quantitative variables
  - No missing data
  - No confounded variables

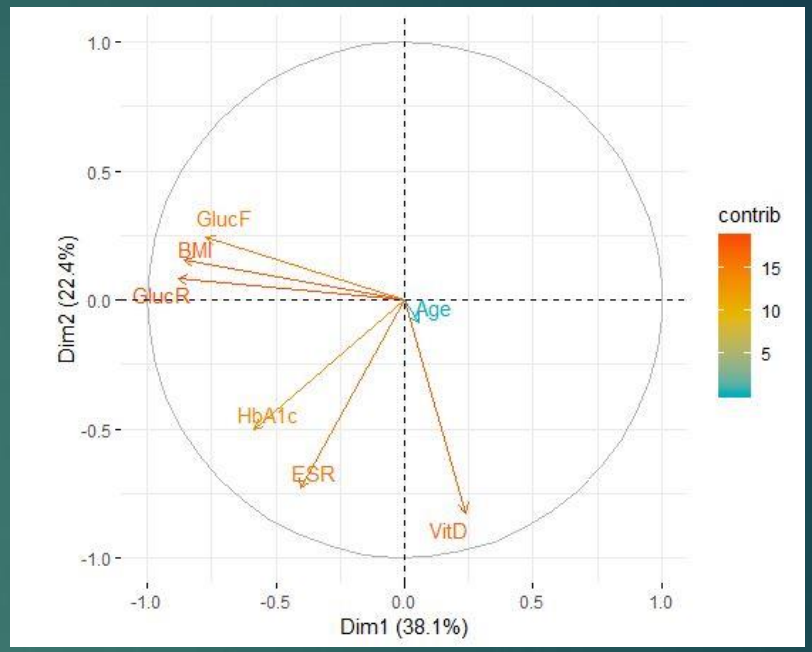
	BMI	ESR	GlucF	GlucR	HbA1c	VitD	Age
1	32.7	26	128	160	6.5	29.2	53.3
2	26.9	11	90	120	2.8	29.7	55.0
3	24.7	13	78	110	2.1	31.0	63.8
4	19.6	11	61	108	2.7	39.4	59.4
5	26.3	10	94	103	1.3	34.3	77.5
6	27.6	9	89	121	3.2	31.1	87.4
7	25.3	10	97	119	2.9	34.1	33.0
8	25.5	10	83	117	3.1	32.0	60.7
9	20.7	22	90	124	4.7	40.6	68.2
10	25.1	11	74	98	5.6	21.5	46.5
11	28.0	17	101	115	4.0	36.7	90.6
12	26.7	17	68	119	3.8	30.3	47.1
13	28.6	24	116	159	5.5	41.1	87.6
14	26.8	24	89	108	3.6	43.5	47.6
15	22.4	7	77	113	1.9	29.0	67.2

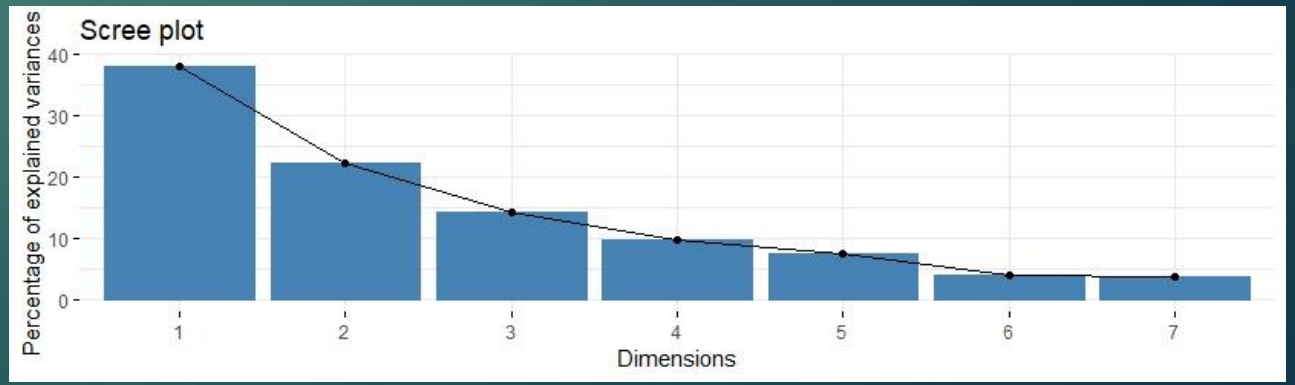
Showing 1 to 15 of 719 entries

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
BMI	-0.52341244	0.12230214	-0.0001441989	0.2609420	-0.27628954	0.72158126	0.214418832
ESR	-0.24841877	-0.58412928	0.1060956418	0.3462488	-0.55241296	-0.38898805	-0.097287182
GlucF	-0.47222419	0.19359747	-0.0085866484	0.4246404	0.51630264	-0.42908092	0.329320952
GlucR	-0.53883849	0.06465371	-0.0500161105	-0.2396698	0.16540226	-0.01825353	-0.786027122
HbA1c	-0.35871625	-0.40044131	-0.0644483141	-0.7046507	0.08002623	-0.04051329	0.449708527
VitD	0.14532995	-0.66097260	0.0430943993	0.2592124	0.56159613	0.37629127	-0.126336473
Age	0.03444981	-0.07026433	-0.9900291105	0.1026466	-0.05275679	-0.01813037	-0.008377289

Principal Components Analysis (PCA)

- Calculate PCA , review loading values, eigenvectors
- More similar variables will occupy the same space (i.e., Random & Fasting Glucose)
- Variables opposite from one another are negatively correlated (i.e., Fib vs PT)





Correspondence Analysis

Use Case:
Identify commonalities among qualitative variables.

Key Assumptions:

- Qualitative variables
- No missing data
- Significant Chi-square (contingency table) result

Contingency Table

	Normal	Obese	OverWt	UnderWt
COMM	143	47	202	0
GOVT	32	12	62	0
OTHER	17	2	33	1
UNINS	66	17	85	0

Chi-square Result

```
> chisq.test(BMICov)  ## Chi-square statistics not appear significant (p=0.38)

        Pearson's Chi-squared test

data:  BMICov
X-squared = 6.4011, df = 6, p-value = 0.3798
```

Correspondence Analysis

- Distribution/distances based on proportions of values in relation to one another.
- Linear algebra is applied to cross-tabulated, scaled data to generated 2D coordinates.
- More similar categories will occupy the same space
- Qualitative data exploration tool, highlights relationship between rows & columns in a table.