Code ▾

Comparing Dog Breeds

Introduction

Let’s explore some features of dog breeds.

In this notebook, I will start by completing the following **Data Manipulation** tasks:

1. Remove unwanted columns and rows
2. Edit strings to remove unwanted characters
3. Convert factor columns to a numeric data type
4. View the data for issues
5. Consider imputing missing values
6. Create a function to assess normality

Next, I will explore relationships in the data via **Data Visualization**. I will create the following:

7. Correlation Matrix
8. Scatterplot: Do size categories make sense?
9. Boxplot: Do larger dogs cost less?
10. Trend Lines: Do larger dogs cost less when controlling for lifespan?
11. Linear Fixed Effects Model
12. Facet_wrap: How does breed category relate to intelligence, genetic ailments, and size?

The data set I will be using is from the “Best in Show” data set collected by David McCandless on https://informationisbeautiful.net/data/ (https://informationisbeautiful.net/data/). The data set is derived from here: https://docs.google.com/spreadsheets/d/1I_HfF5EaN-QgnLc2UYdCc7L2CVrk0p3VdGB1godOyhk/edit#gid=20 (https://docs.google.com/spreadsheets/d/1I_HfF5EaN-QgnLc2UYdCc7L2CVrk0p3VdGB1godOyhk/edit#gid=20)

Data Manipulation

I will start by loading the csv file from my computer.

Hide

Hide

```
library("dplyr")
# Import Data
dog <- read.csv("C:\\Users\\Public\\Documents\\best_in_show.csv")
head(dog)
```

	Dog.breed <fctr>	X <lgl>	category <fctr>	… <lgl>	datadog.score <dbl>	▶
1	Additional info	NA	American Kennel Club group	NA	NA	
2	Border Collie	NA	herding	NA	3.64	
3	Border Terrier	NA	terrier	NA	3.61	

	Dog.breed <fctr>	X <lgl>	category <fctr>	… <lgl>	datadog.score <dbl> ▸
4	Brittany	NA	sporting	NA	3.54
5	Cairn Terrier	NA	terrier	NA	3.53
6	Welsh Springer Spaniel	NA	sporting	NA	3.34

6 rows | 1-6 of 69 columns

1. Remove unwanted columns and rows

Each row is a distinct dog breed. When scrolling through the columns, we can see that there are several we don’t need or understand.

Let’s only keep the columns of interest, and let’s rename them. I will also remove the first row, as it is only contains notes about each column.

Hide

Hide

```
dog <- select(dog, c(Dog.breed, category, POPULARITY.IN.US, LIFETIME.COST...,
                     X1.INTELLIGENCE..TRAINABILITY..ranking, X2.LONGEVITY, X3.NO..OF.GENETIC.AIL
MENTS,
                     X4a.average.purchase.price..US., size.category, weight..kg.,
                     shoulder.height..cm., intelligence.category))
dog <- rename(dog, c(Breed = Dog.breed, Breed_category = category, Popularity_rank = POPULARITY.
IN.US,
               Lifetime_cost = LIFETIME.COST..., Intelligence_rank = X1.INTELLIGENCE..TRAINABILI
TY..ranking,
               Lifespan = X2.LONGEVITY, Genetic_ailments = X3.NO..OF.GENETIC.AILMENTS,
               Purchase_price = X4a.average.purchase.price..US., Size_category = size.category,
               Weight = weight..kg., Height = shoulder.height..cm., Intelligence_category = inte
lligence.category
               ))
dog <- dog[-1, ]
head(dog, 10)
```

	Breed <fctr>	Breed_category <fctr>	Popularity_rank <fctr>	Lifetime_cost <fctr>	Intelligence_ra <fctr>
2	Border Collie	herding	45	$20,143	1
3	Border Terrier	terrier	80	$22,638	30
4	Brittany	sporting	30	$22,589	19
5	Cairn Terrier	terrier	59	$21,992	35
6	Welsh Springer Spaniel	sporting	130	$20,224	31
7	English Cocker Spaniel	sporting	63	$18,993	18

	Breed <fctr>	Breed_category <fctr>	Popularity_rank <fctr>	Lifetime_cost <fctr>	Intelligence_ra <fctr>
8	Cocker Spaniel	sporting	27	$24,330	20
9	Papillon	toy	38	$21,001	8
10	Australian Cattle Dog	herding	60	$20,395	10
11	Shetland Sheepdog	herding	20	$21,006	6

1-10 of 10 rows | 1-7 of 12 columns

2. Edit strings & 3. Convert columns to numeric data types

It looks like all of the columns are factors, but we want some of them to be numeric.

Columns where the strings already look like numbers are good to go, but the columns with dollar amounts need to have the “$” and “,” characters replaced with nothing.

Hide

Hide

```
ready_columns <- list(3, 5, 6, 7, 10, 11)
price_columns <- list(4, 8)
for (i in 3:11){
  if (is.element(i, ready_columns)) {
    dog[,i] <- as.numeric(as.character(dog[,i]))

  } else if (is.element(i, price_columns)) {
    # replace first character ("$") and/or all commas with nothing
    dog[,i] <- as.numeric(gsub("^.|,","", dog[,i]))
  }
}
```

```
NAs introduced by coercionNAs introduced by coercionNAs introduced by coercionNAs introduced by
coercionNAs introduced by coercionNAs introduced by coercionNAs introduced by coercion
```

Hide

Hide

```
head(dog)
```

	Breed <fctr>	Breed_category <fctr>	Popularity_rank <dbl>	Lifetime_cost <dbl>	Intelligence_
2	Border Collie	herding	45	20143	
3	Border Terrier	terrier	80	22638	

	Breed <fctr>	Breed_category <fctr>	Popularity_rank <dbl>	Lifetime_cost <dbl>	Intelligence_
4	Brittany	sporting	30	22589	
5	Cairn Terrier	terrier	59	21992	
6	Welsh Springer Spaniel	sporting	130	20224	
7	English Cocker Spaniel	sporting	63	18993	

6 rows | 1-7 of 12 columns

4. View the data

Now let’s look at a summary of the data.

Hide

Hide

```
# View the data frame
summary(dog)
```

```
                        Breed          Breed_category Popularity_rank  Lifetime_cost   Intellige
nce_rank
 Affenpinscher             :  1   sporting    :28     Min.   :  1.00   Min.   :12653   Min.   :
1.00
 Afghan Hound              :  1   terrier     :28     1st Qu.: 43.75   1st Qu.:17817   1st Qu.:2
7.00
 Airedale Terrier          :  1   working     :27     Median : 87.50   Median :20087   Median :4
2.00
 Akita                     :  1   hound       :26     Mean   : 87.12   Mean   :19820   Mean   :4
0.92
 Alaskan Malamute          :  1   herding     :25     3rd Qu.:130.25   3rd Qu.:21798   3rd Qu.:5
4.25
 American English Coonhound:  1   non-sporting:19     Max.   :173.00   Max.   :26686   Max.   :8
0.00
 (Other)                   :168   (Other)     :21     NA's   :2        NA's   :83      NA's   :4
2
    Lifespan     Genetic_ailments Purchase_price   Size_category     Weight           Height
 Min.   : 6.29   Min.   :0.000    Min.   : 283.0         : 2     Min.   : 2.00   Min.   :13.00
 1st Qu.: 9.70   1st Qu.:0.000    1st Qu.: 587.2   large :54     1st Qu.: 8.00   1st Qu.:36.00
 Median :11.29   Median :1.000    Median : 795.0   medium:60     Median :16.00   Median :48.00
 Mean   :10.96   Mean   :1.216    Mean   : 876.8   small :58     Mean   :20.35   Mean   :48.52
 3rd Qu.:12.37   3rd Qu.:2.000    3rd Qu.:1042.2                 3rd Qu.:28.00   3rd Qu.:61.50
 Max.   :16.50   Max.   :9.000    Max.   :3460.0                 Max.   :79.00   Max.   :81.00
 NA's   :39      NA's   :26       NA's   :28                     NA's   :88      NA's   :15
   Intelligence_category
 no data      :40
 Average      :39
 Above average:29
 Fair         :22
 Excellent    :21
 Lowest       :11
 (Other)      :12
```

5. Consider imputing missing values

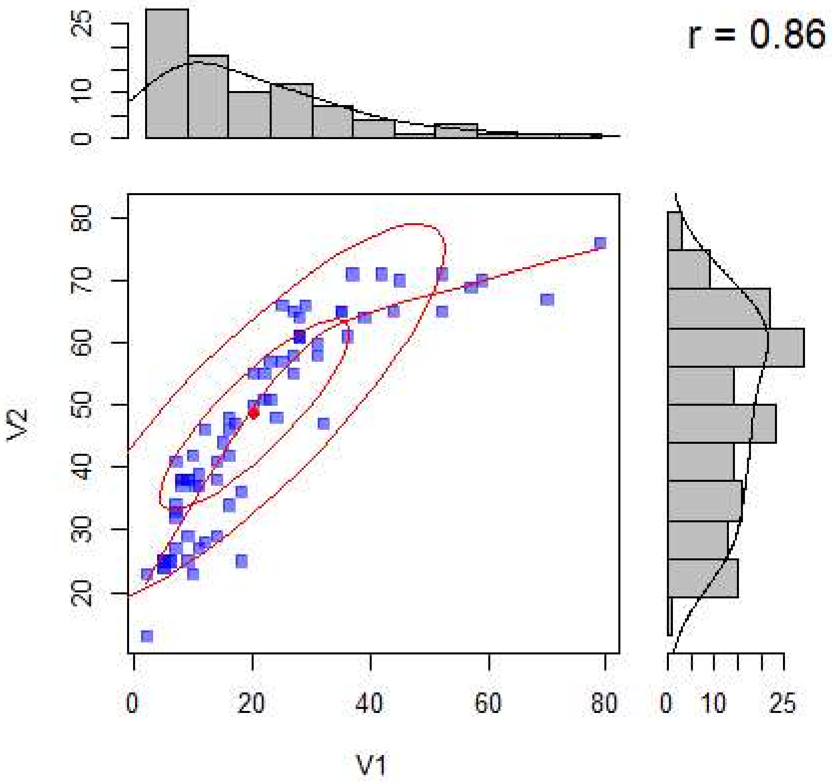
Knowing that there is one entry for every breed, we can see that there are 174 breeds included in this data set.

It seems like there are a lot of missing values for Weight. Could weight and height be highly correlated enough to impute the missing data?

Hide

Hide

```
library(psych) #for the scatterplot/histogram
# View the relationship
scatter.hist(dog$Weight, dog$Height)
```



Hide

Hide

```
# Print the correlation coefficient:
sprintf("Correlation coefficient: %f", cor(dog$Weight, dog$Height, use="complete.obs"))
```

```
[1] "Correlation coefficient: 0.858692"
```

Hide

Hide

```
# Test
cor_sp <- cor.test(dog$Weight, dog$Height, method = "spearman", use = "complete.obs", exact=FALS
E)
cor_kn <- cor.test(dog$Weight, dog$Height, method = "kendall", use = "complete.obs", exact=FALSE
)
sprintf("Spearman test: p = %s", cor_sp$p.value)
```

```
[1] "Spearman test: p = 6.38615292613551e-31"
```

Hide

Hide

```
sprintf("Kendall test: p = %s", cor_kn$p.value)
```

```
[1] "Kendall test: p = 5.23260420647559e-22"
```

Given the very high, statistically significant correlation between height and weight, I would say that height could be used to impute the missing values in weight. However, it is not good practice to impute when roughly 20% or more of the data is missing, and there are currently 88 / 174 missing values, or about 50% missing. So, I will forgo imputing the missing values this time.

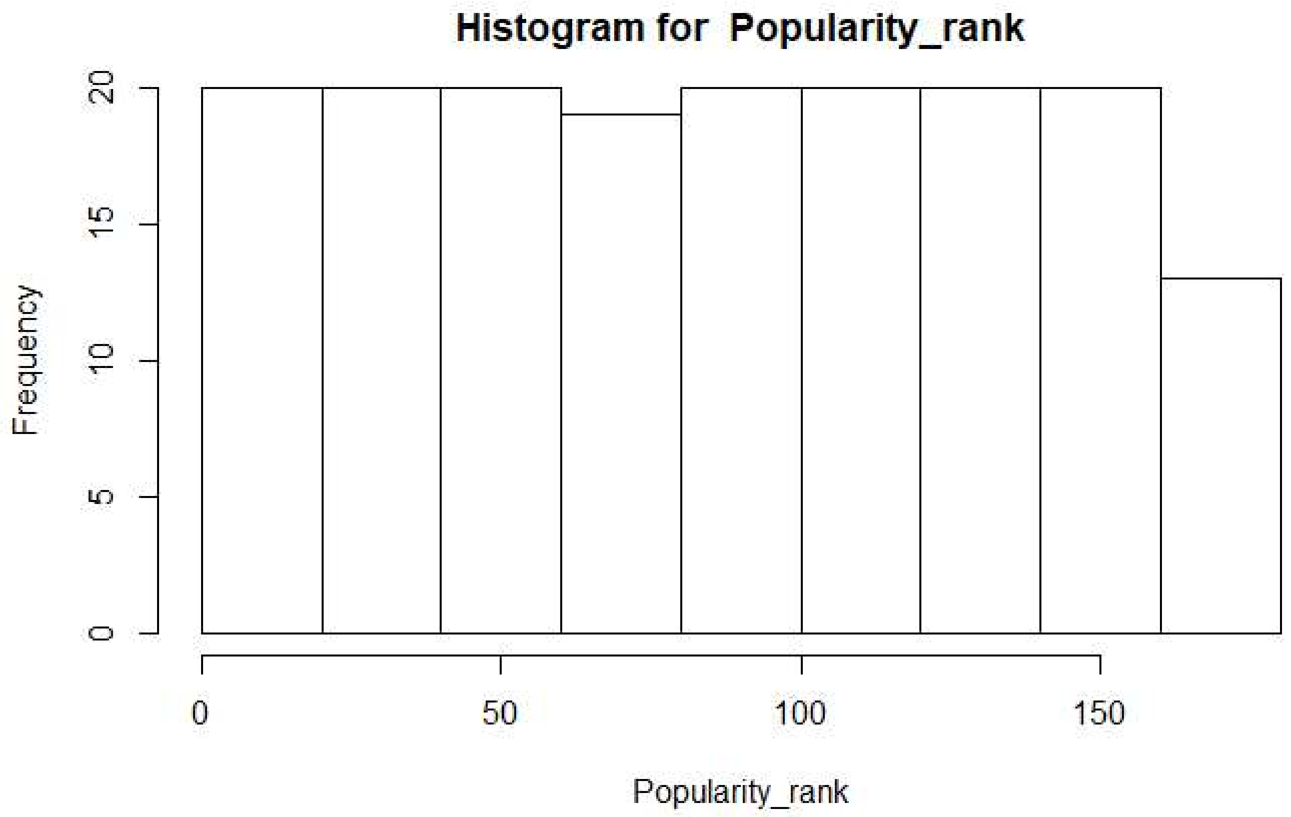
6. Create a function to assess normality

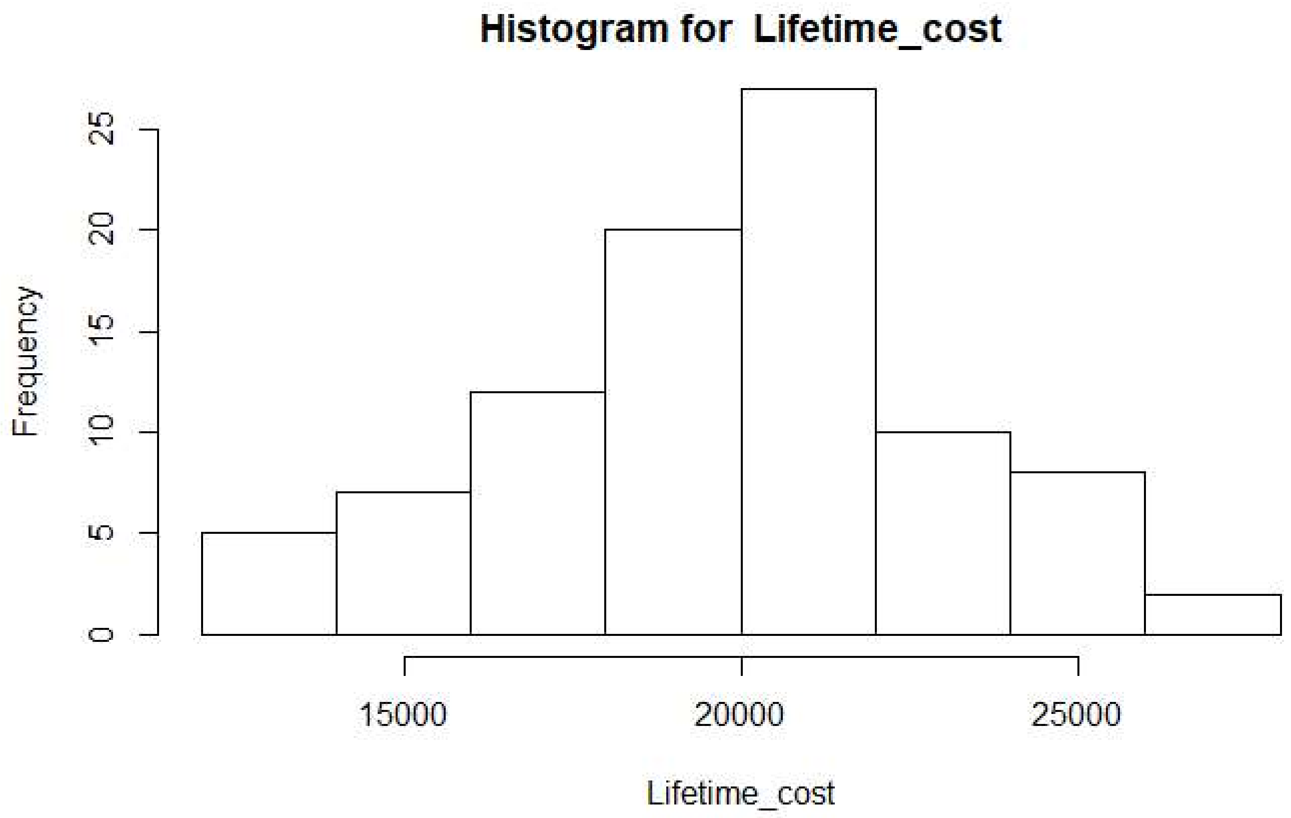
Are there other interesting relationships we can pull from the data? First, let’s assess for normality to inform what type of statistical test to use later on.

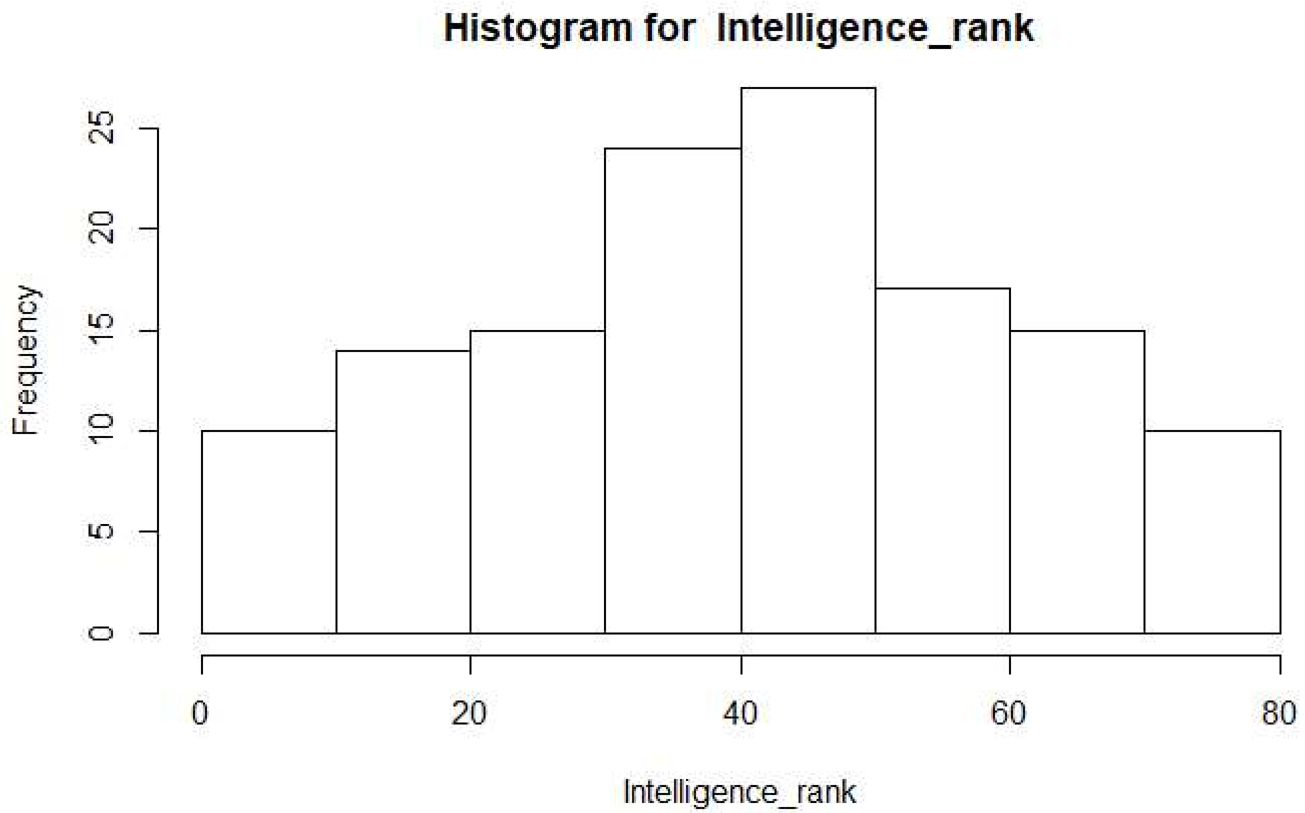
Hide

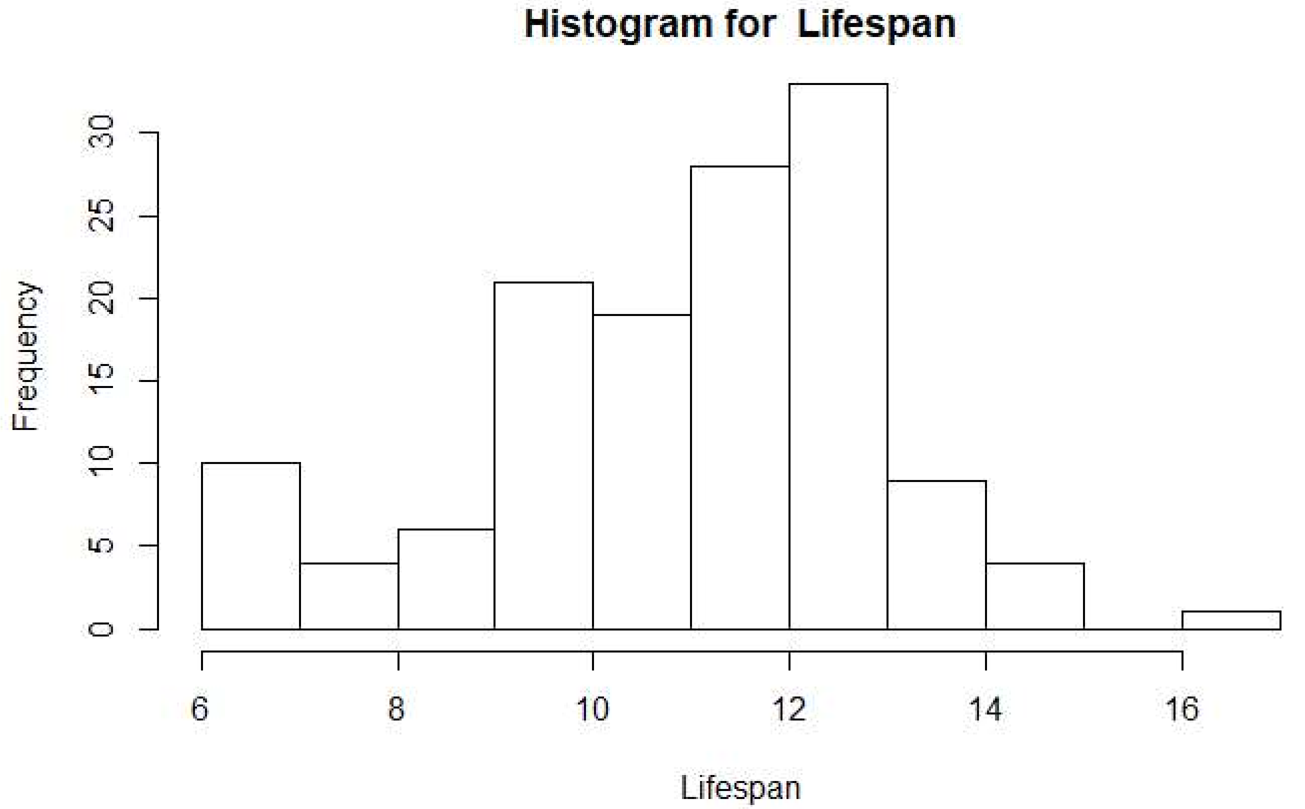
Hide

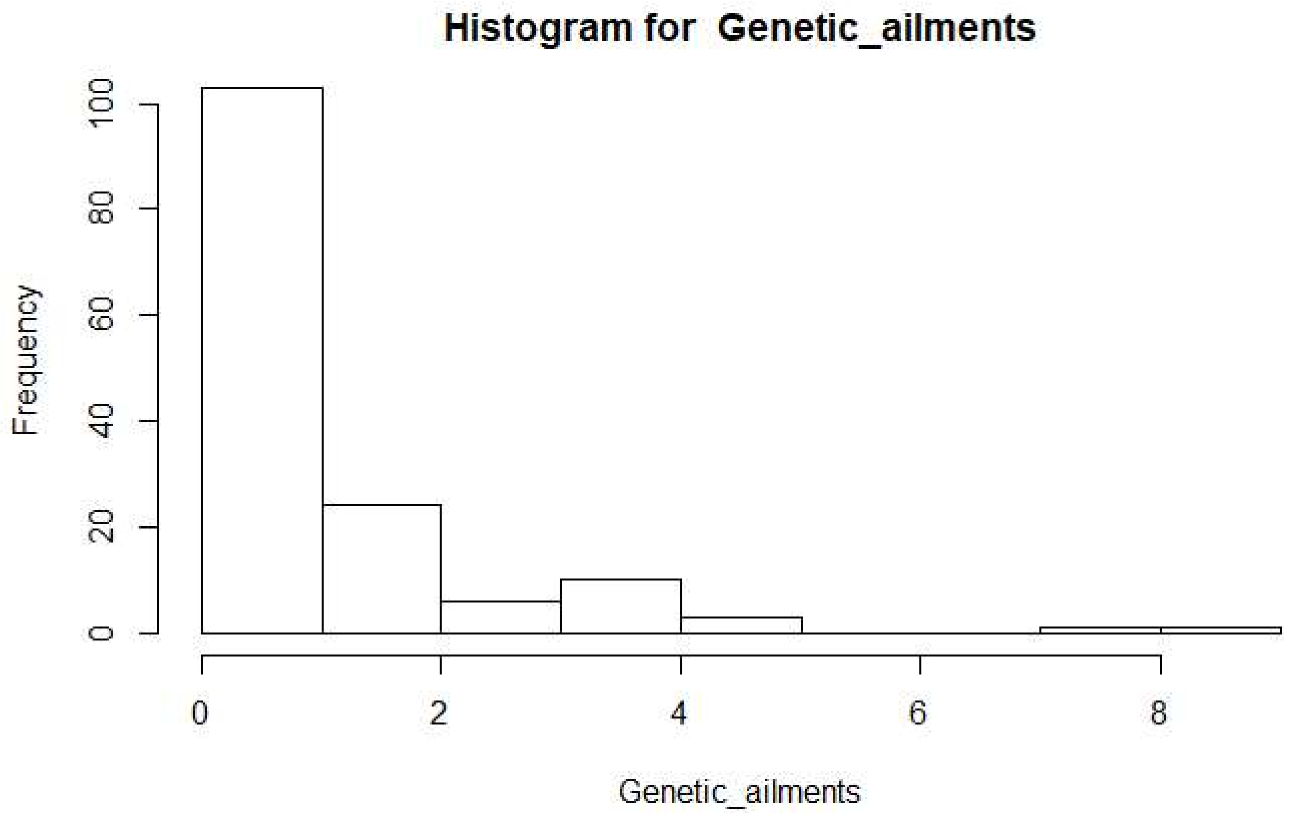
```
# Create a function that checks for normality on all numerical variables.
normality <- function(input_df) {
############### VISUALS #############
  for (s in 1:length(input_df)){
    cname = colnames(input_df[s])
    ## 1. See if bell-shaped:
    hist(input_df[,s],
         main=paste("Histogram for ", cname),
         xlab = cname)

  }
############### TESTS ###############
  library(moments)
  ## 1. Shapiro-Wilk’s test.
  ## Null hypothesis: the distribution is normal.
  ## If the test is significant, then the distribution is non-normal.

  ## 2. Skewness Test.
  ## A skew between -0.5 and 0.5 is normal, up to abs(1) is moderate, and above is probably more severe.
  ## A negative skew is left-skewed, or the mean is less than the median.
  dog_norm <- setNames(data.frame(matrix(ncol = 6, nrow = length(input_df)), stringsAsFactors = FALSE),
                     c("variable", "shapiro_stat", "shapiro_p_value",
                       "shapiro_normal", "skew", "skewness_strength"))
  for (s in 1:length(input_df)){
    result <- shapiro.test(input_df[,s])

    dog_norm[s,1] <- colnames(input_df[s])
    dog_norm[s,2] <- result$statistic
    dog_norm[s,3] <- result$p.value
    if (result$p.value < 0.05){
      dog_norm[s,4] <- "non-paramteric"
    } else {
      dog_norm[s,4] <- "normal"
    }
    dog_norm[s,5] <- skewness(input_df[s], na.rm = TRUE)

    if (abs(skewness(input_df[s], na.rm = TRUE)) < 0.5){
      dog_norm[s,6] <- "normal"
    } else if (abs(skewness(input_df[s], na.rm = TRUE)) < 1.0) {
      dog_norm[s,6] <- "moderate_skew"
    } else {
      dog_norm[s,6] <- "severe_skew"
    }
  }

  return(dog_norm)
}
# Call the function!
dog2 <- normality(dog[c(3:8,10,11)])
```

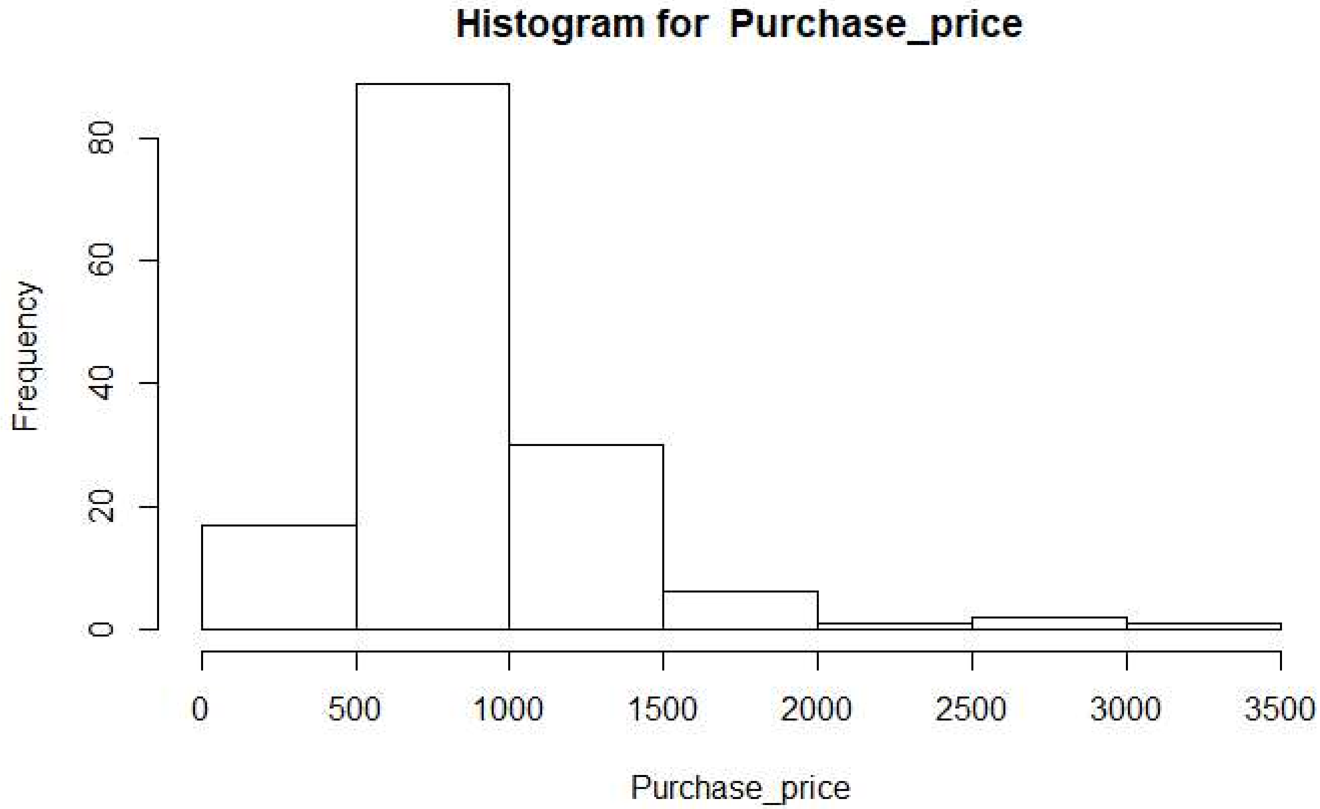

Histogram for Popularity_rank
Frequency
0
5
10
15
20
0
50
100
150
Popularity_rank

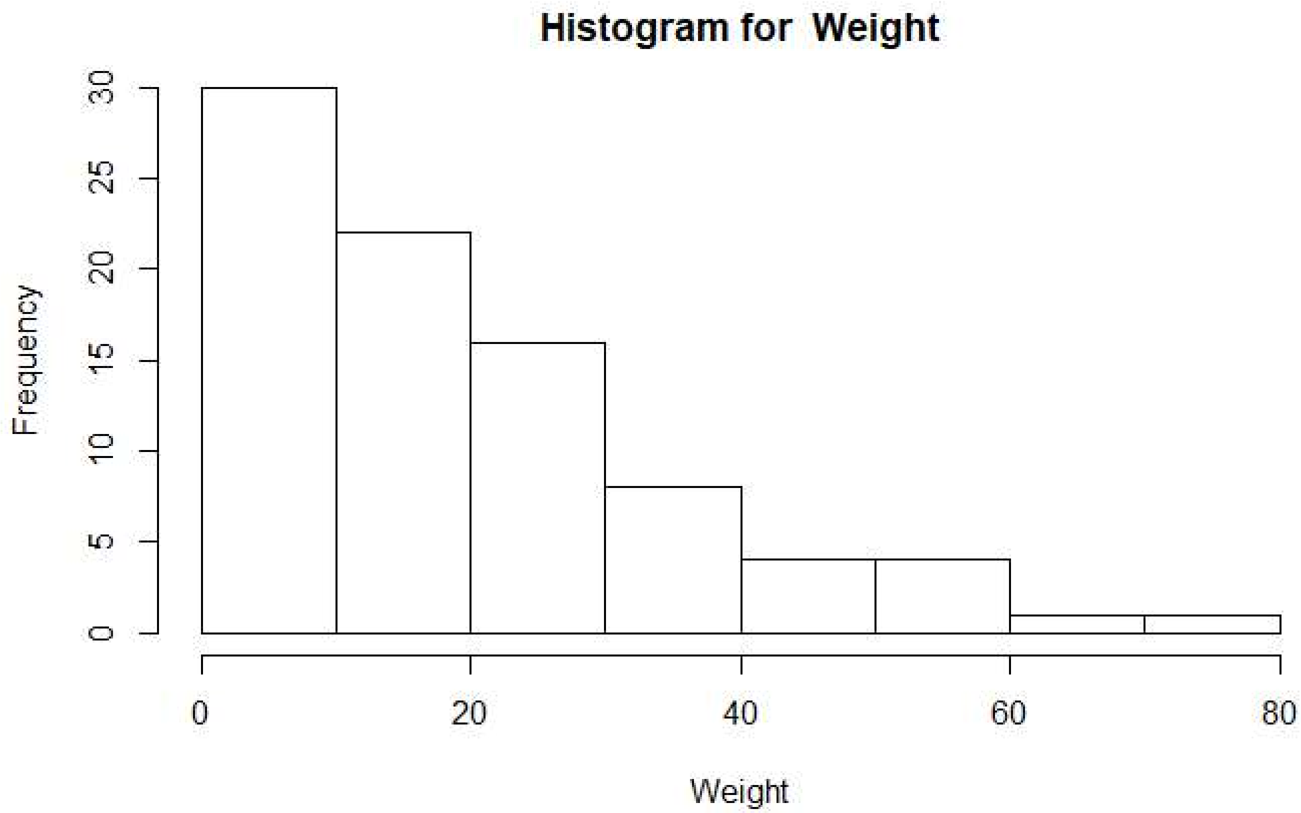

Histogram for Lifetime_cost
Frequency
0
5
10
15
20
25
15000
20000
25000
Lifetime_cost

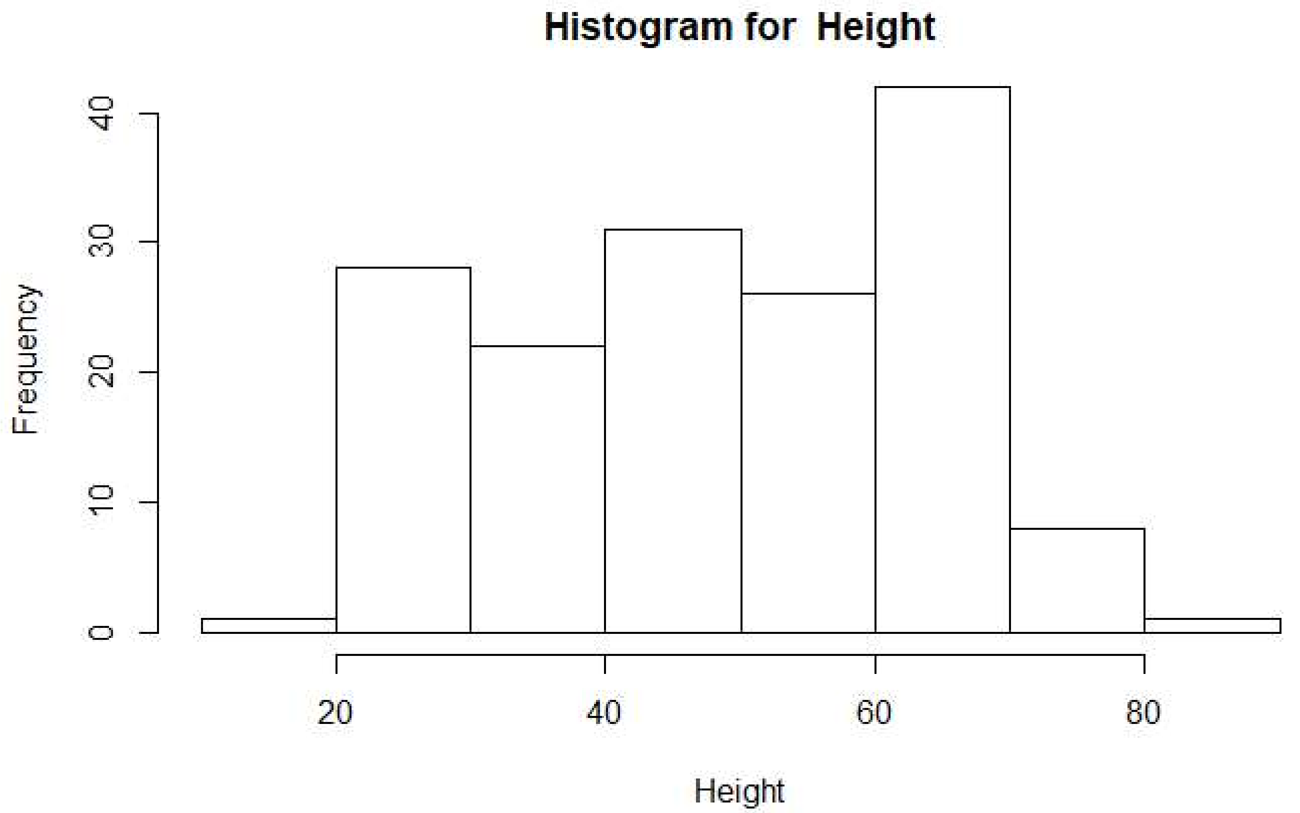

Histogram for Intelligence_rank
Frequency
0
5
10
15
20
25
0
20
40
60
80
Intelligence_rank


Histogram for Lifespan
Frequency
0
5
10
15
20
25
30
6
8
10
12
14
16
Lifespan


Histogram for Genetic_ailments
Frequency
0
20
40
60
80
100
0
2
4
6
8
Genetic_ailments


Histogram for Purchase_price
Frequency
0
20
40
60
80
0
500
1000
1500
2000
2500
3000
3500
Purchase_price





Hide

Hide

```
head(dog2,10)
```

	variable <chr>	shapiro_stat <dbl>	shapiro_p_value <dbl>	shapiro_normal <chr>	skew <dbl>	skewnes <chr>
1	Popularity_rank	0.9538875	2.025039e-05	non-paramteric	-0.006887718	normal
2	Lifetime_cost	0.9897098	7.029169e-01	normal	-0.060836675	normal
3	Intelligence_rank	0.9832419	1.042006e-01	normal	-0.059317115	normal
4	Lifespan	0.9691551	3.702723e-03	non-paramteric	-0.415710639	normal
5	Genetic_ailments	0.7531262	1.716435e-14	non-paramteric	2.012404075	severe_s
6	Purchase_price	0.7973670	6.147461e-13	non-paramteric	2.438912780	severe_s
7	Weight	0.8724653	4.700155e-07	non-paramteric	1.356831337	severe_s
8	Height	0.9557300	6.121421e-05	non-paramteric	-0.172397191	normal

8 rows

Visually and based on the skewness, it seems like many factors are normally distributed. However, given the results of the Shapiro-Wilk's test, non-parametric testing might be better suited to all but "Lifetime_cost" and "Intelligence_rank."

Data Visualization

7. Correlation Matrix

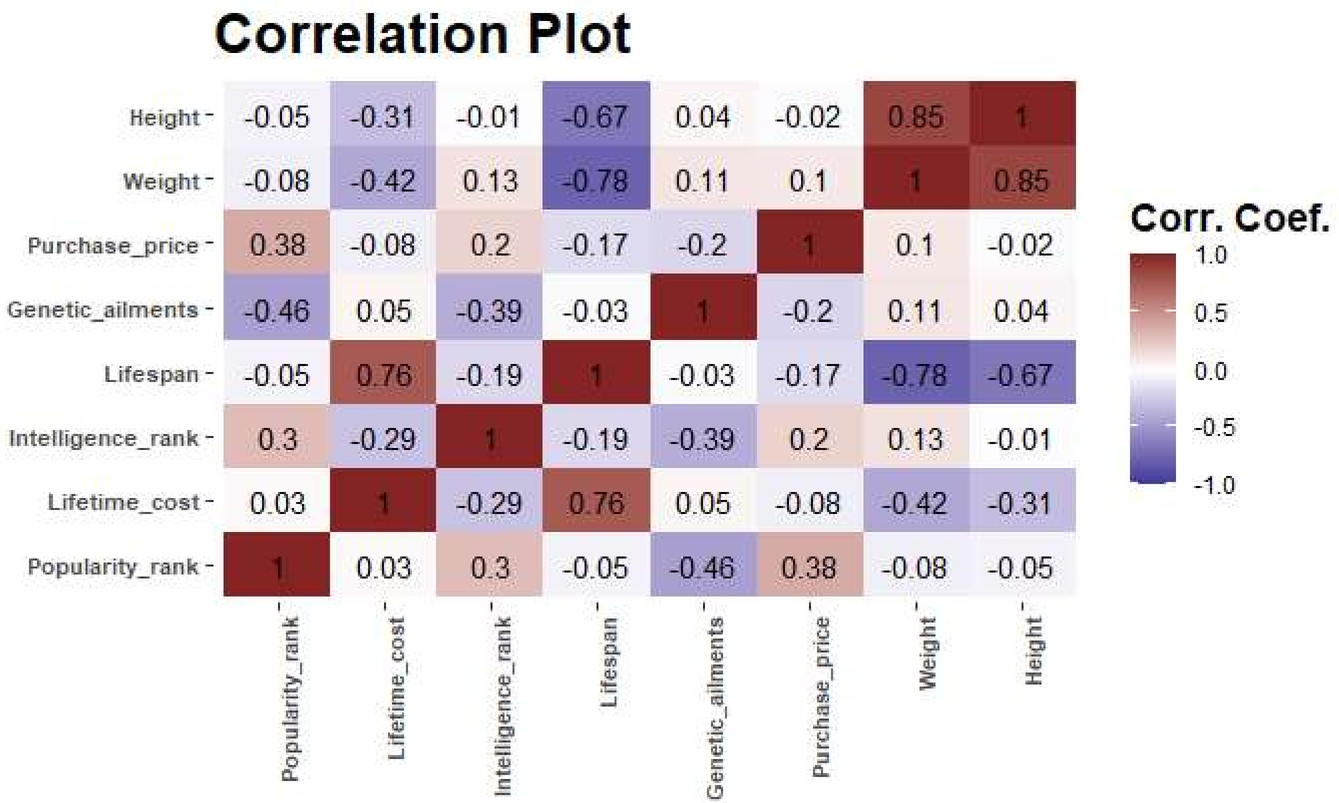
Now, let's see how these factors are related.

Hide

Hide

```
library(reshape2)
library(ggplot2)
library(scales) # for muted function
corr_mat <- cor(dog[c(3:8,10,11)],
                dog[c(3:8,10,11)],
                use = "complete.obs")
co=melt(corr_mat)
ggplot(co, aes(Var1, Var2)) +
  geom_tile(aes(fill = value)) + # background colours are mapped according to the value column
  geom_text(aes(fill = value, label = round(value, 2))) + # write the values
  scale_fill_gradient2(low = muted("midnightblue"),
                       mid = "white",
                       high = muted("darkred"),
                       midpoint = 0,
                       limits = c(-1,1)) + # determine the colour
  theme(panel.grid.major.x=element_blank(), #no gridlines
        panel.grid.minor.x=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.minor.y=element_blank(),
        panel.background=element_rect(fill="white"), # background=white
        axis.text.x = element_text(angle=90, hjust = 1,vjust=1,size = 8,face = "bold"),
        plot.title = element_text(size=20,face="bold"),
        axis.text.y = element_text(size = 8,face = "bold")) +
  ggtitle("Correlation Plot") +
  theme(legend.title=element_text(face="bold", size=14)) +
  scale_x_discrete(name="") +
  scale_y_discrete(name="") +
  labs(fill="Corr. Coef.")
```

```
Ignoring unknown aesthetics: fill
```



	Popularity_rank	Lifetime_cost	Intelligence_rank	Lifespan	Genetic_ailments	Purchase_price	Weight	Height
Height	-0.05	-0.31	-0.01	-0.67	0.04	-0.02	0.85	1
Weight	-0.08	-0.42	0.13	-0.78	0.11	0.1	1	0.85
Purchase_price	0.38	-0.08	0.2	-0.17	-0.2	1	0.1	-0.02
Genetic_ailments	-0.46	0.05	-0.39	-0.03	1	-0.2	0.11	0.04
Lifespan	-0.05	0.76	-0.19	1	-0.03	-0.17	-0.78	-0.67
Intelligence_rank	0.3	-0.29	1	-0.19	-0.39	0.2	0.13	-0.01
Lifetime_cost	0.03	1	-0.29	0.76	0.05	-0.08	-0.42	-0.31
Popularity_rank	1	0.03	0.3	-0.05	-0.46	0.38	-0.08	-0.05

A lot of interesting correlations pop out.

1. What is correlated to the **SIZE** (height and weight) of the dog breed?

- It seems that larger dog breeds are associated with shorter lifespans (r=-0.78 for weight, r=-0.67 for height) and smaller lifetime costs (r=-0.42 and -0.31 respectively).
- This corresponds with the high correlation between lifespan and lifetime costs (r=0.76): it is likely that larger breeds cost less due to their shortened lifespan.

2. What is correlated to the **INTELLIGENCE** of the dog breed?

- It appears that the more intelligent dog breeds are:
- more popular (r=0.3),
- have fewer genetic ailments (r=-0.39),
- are potentially heavier than other breeds of the same height (r=0.13 for weight while r=-0.01 for height), and
- are more expensive to purchase (r=0.2) but cheaper over a lifetime (r=-0.29).

3. What makes a dog breed **POPULAR** in the US?

- The most popular dog breeds are associated with:
- greater intelligence (r=0.3),
- fewer genetic ailments (r=-0.46), and
- a higher purchase price (r= 0.38).

8. Scatterplot: Do size categories make sense?

Let’s take a deeper look at the relationship between breed lifespan, lifetime cost, and size. The Size_category factor will be used instead of height or weight for this visualization. To prepare that column for visualization, **the missing values will need to be replaced by “NA.”** Let’s do that for all three factors.

Hide

Hide

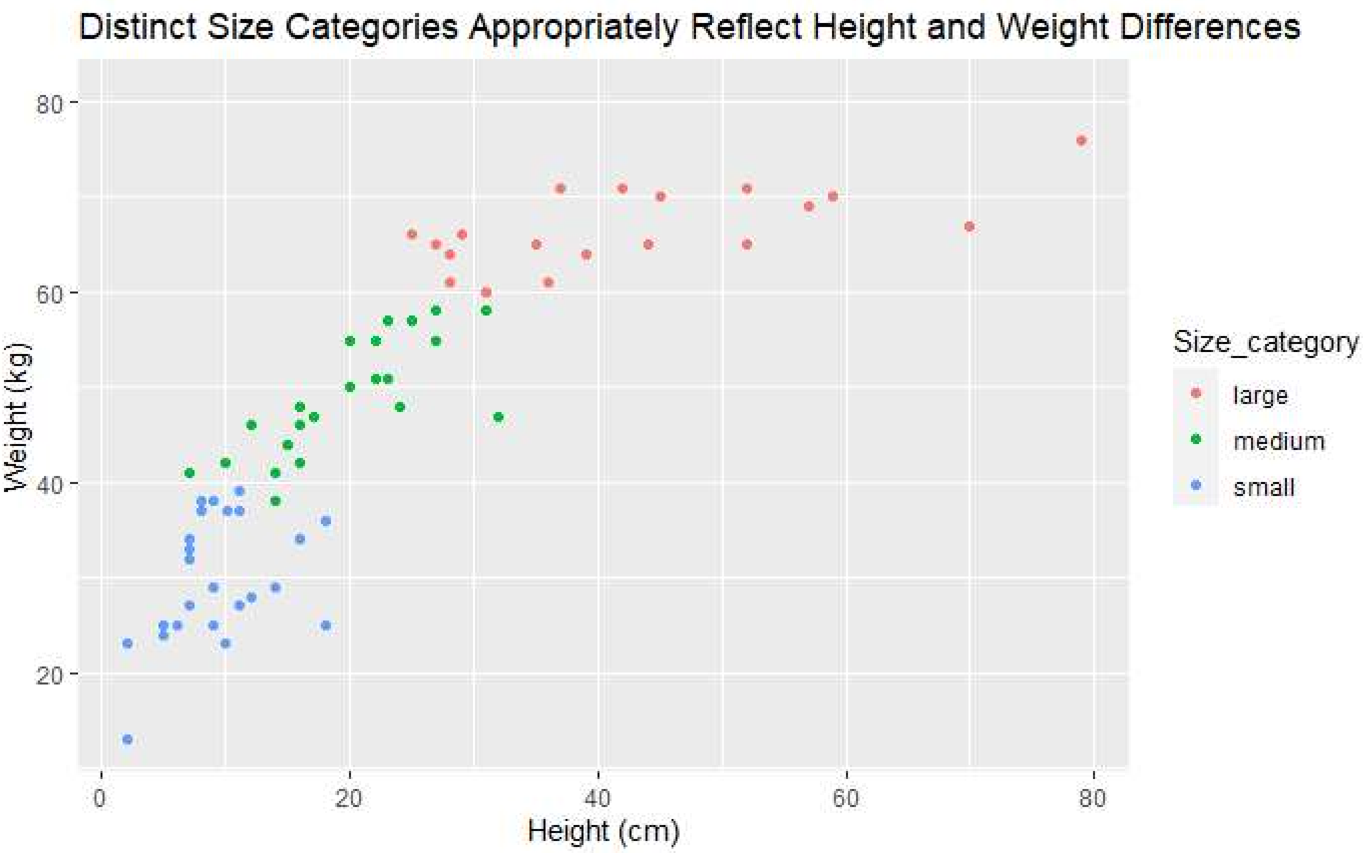
```
dog$Size_category <- dplyr::recode(dog$Size_category, large='large',
                                    medium='medium', small='small',
                                    .default=NA_character_)
dog$Breed_category <- dplyr::recode(dog$Breed_category, .herding='herding', hound='hound',
                                     `non-sporting`='non-sporting', sporting='sporting',
                                     terrier='terrier', toy='toy', working='working',
                                     .default=NA_character_, `American Kennel Club Group`=NA_cha
racter_)
dog$Intelligence_category <- dplyr::recode(dog$Intelligence_category, Brightest='Brightest',
                                            Excellent='Excellent', `Above average`='Above averag
e',
                                            Average='Average', Fair='Fair', Lowest='Lowest',
                                            `no data`=NA_character_, .default=NA_character_)
```

Now, let’s check that the Size_category appropriately categorizes breeds based on height and weight.

Hide

Hide

```
dog %>%
  drop_na(Size_category) %>%
ggplot() +
  aes(x = Weight,
      y = Height,
      color=Size_category) +
  geom_point() +
  labs(title = "Distinct Size Categories Appropriately Reflect Height and Weight Differences",
       x = "Height (cm)",
       y = "Weight (kg)")
```



Hide

Hide

```
summary(dog$Size_category)
```

```
 large medium  small   NA's 
    54     60     58      2 
```

Yes. Each category has roughly the same number of breeds (between 54 and 60), there are very few missing values (2), and the categories seem to follow coherent distinctions based mostly on height.

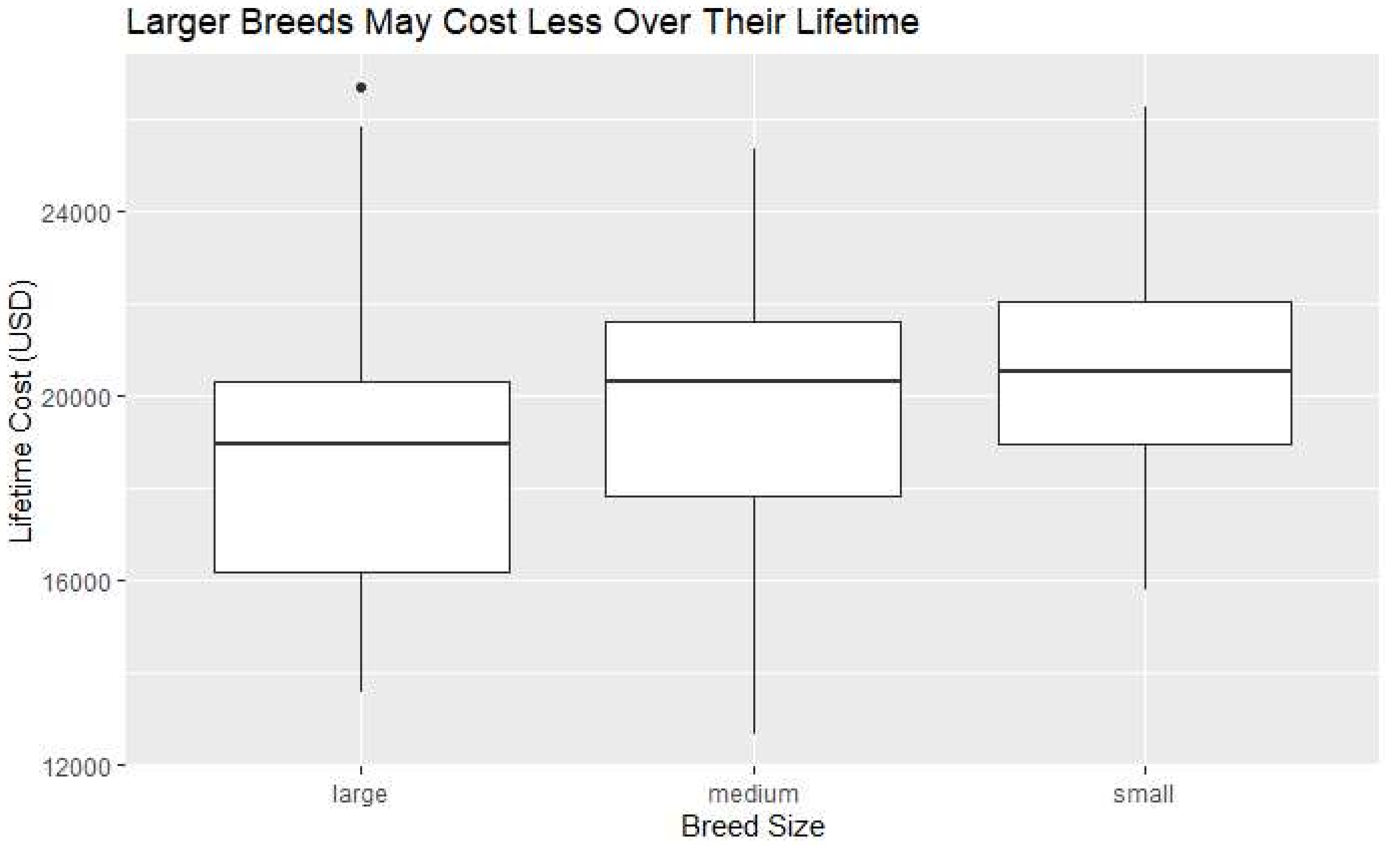
9. Boxplot: Do larger dogs cost less?

Next, let’s quickly see whether larger breeds have a lower lifetime cost, as previously indicated by the correlation matrix.

Hide

Hide

```
dog %>%
  drop_na(Size_category) %>%
ggplot() +
  aes(x = Size_category,
      y = Lifetime_cost) +
  geom_boxplot() +
  labs(title = "Larger Breeds May Cost Less Over Their Lifetime",
       x = "Breed Size",
       y = "Lifetime Cost (USD)")
```



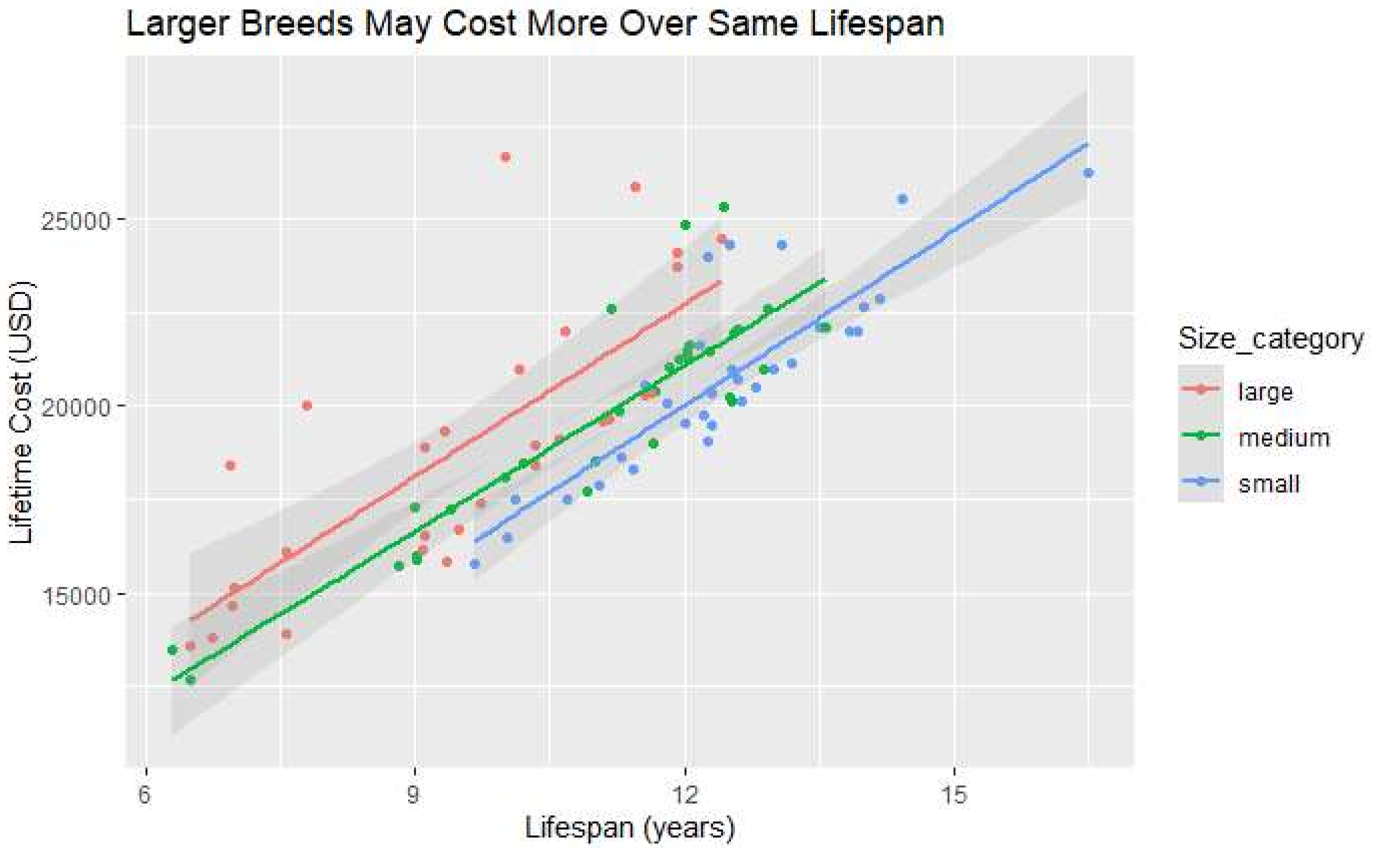
It would appear that larger breeds do cost less. However, the length of a dog’s lifetime will likely affect its total cost over its lifetime. Do larger breeds cost less because they have shorter lifespans, or is it due to some other inherent feature of being a larger breed? If we compare breeds with equal lifespans, will size impact the lifetime cost?

10. Trend Lines: Do larger dogs cost less when controlling for lifespan?

Hide

Hide

```
library(tidyr)
dog %>%
  drop_na(Size_category) %>%
ggplot() +
  aes(x = Lifespan,
      y = Lifetime_cost, color = Size_category) +
  geom_point() +
  geom_smooth(method=lm, level=0.95, alpha=0.2) +
  labs(title = "Larger Breeds May Cost More Over Same Lifespan",
       x = "Lifespan (years)",
       y = "Lifetime Cost (USD)")
```



We can see clearly that, when aggregating across all dog sizes, there is a strong positive relationship between lifespan and lifetime cost. However, once we have separated this data based on dog size, we see that larger dog breeds tend to cost more than smaller breeds when their lifespans are the same. This is suggested by the difference in the trend lines: the gray bands surrounding each linear trend line represent the 95% confidence intervals. Since the bands for the large and small breed trend lines do not overlap, this suggests that there is likely a difference between the two groups.

Let’s quickly test the main effects of size and lifespan with a linear fixed effects model.

11. Linear Fixed Effects Model

Hide

Hide

```
library(lme4)
form <- formula(Lifetime_cost ~ Lifespan * Size_category)
mod_dog = lm(formula=form,data=dog)
anova(mod_dog)
```

```
Analysis of Variance Table

Response: Lifetime_cost
                       Df    Sum Sq   Mean Sq  F value    Pr(>F)    
Lifespan                1 543028416 543028416 183.2767 < 2.2e-16 ***
Size_category           2  71127040  35563520  12.0030 2.563e-05 ***
Lifespan:Size_category  2    280818    140409   0.0474    0.9537    
Residuals              85 251845542   2962889                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It appears that there is a main effect of lifespan and breed size, as denoted by the strong correlations seen in the correlation matrix previously, but there is no interaction effect.

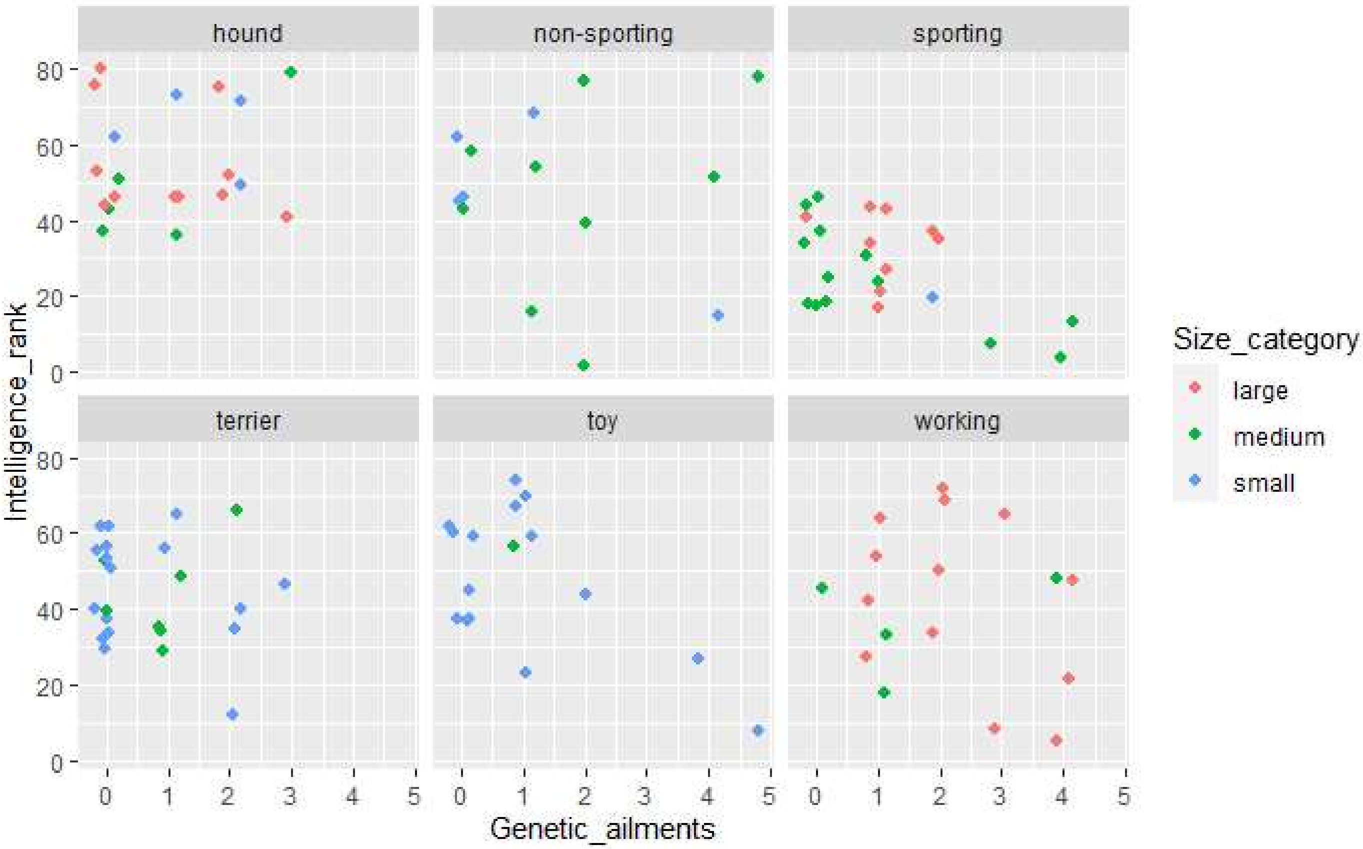
12. Facet_wrap: How does breed category relate to intelligence, genetic ailments, and size?

Next, let’s see whether the breed category has any impact on intelligence, genetic ailments, or the size of the dog.

Hide

Hide

```
dog %>%
  drop_na(Breed_category) %>%
ggplot() +
  aes(x = Genetic_ailments,
      y = Intelligence_rank, color = Size_category) +
    geom_jitter(width=0.2, size=1.8) +
  facet_wrap(~ Breed_category)
```



From this set of plots, we can see:

1. **Intelligence:** Hounds tend to rank highest in intelligence, and sporting breeds tend to rank lowest.
2. **Genetic ailments:** Working breeds seem to have a higher number of genetic ailments than other breeds on average.
3. **Size:** Working and sporting breeds tend to be larger, especially when compared to terrier and toy breeds.

Conclusion

This brief data exploration revealed strong associations between dog breed size and other factors. There was an interesting trend towards larger dogs costing more over their lifetime on average compared to small dog breeds of the same average lifespan, but an additional analysis would have to be conducted to verify this. One way to pursue this hypothesis would be to eliminate lifespan outliers and only look at data points where all three sizes are represented over each included lifespan length, such as in the 9-13 year range. Alternatively, an analysis of covariance (ANCOVA) could be run to determine whether there is a difference in the slope or intercept of the trend lines seen in the plot; this would confirm whether larger dog breeds cost *more* when lifespan is accounted for.

In the future, I plan on importing a few more data sets and merging them with the present one. I think it would be interesting to see what kinds of breeds are registered by dog owners in New York City and Seattle. Are there any differences in the types of dogs these two opposite-coast cities prefer? To be continued.