Exploratory Data Analysis Report

Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

Overview of the data

Descriptions <chr>	**Value** <chr>
Sample size (nrow)	397
No. of variables (ncol)	9
No. of numeric/interger variables	7
No. of factor variables	0
No. of text variables	2
No. of logical variables	0

No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	100% (9)

1-10 of 13 rows Previous **1** 2 Next

Structure of the data

Ind...	Variable_Name	Variable_Type	Sampl...	Missing_Count	Per_of_Missing	No_of_
<dbl>	<chr>	<chr>	<int>	<int>	<dbl>	
1	mpg	numeric	397	0	0	
2	cylinders	integer	397	0	0	
3	displacement	numeric	397	0	0	
4	horsepower	character	397	0	0	
5	weight	integer	397	0	0	
6	acceleration	numeric	397	0	0	
7	year	integer	397	0	0	
8	origin	integer	397	0	0	
9	name	character	397	0	0	

9 rows

Target variable

Summary of continuous dependent variable

1. Variable name - **mpg**
2. Variable description - ****

2. Summary of numerical variables

Summary statistics when dependent variable is Continuous **mpg**.

```
ExpNumStat(data,by="A",gp=Target,Qnt=seq(0,1,0.1),MesofShape=2,Outlier=TRUE,round=2)
```

Vname	Group	Note	TN	n...	nZero	n...	NegInf	PosInf	NA_Value
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

acceleration	mpg	Cor b/w mpg	397	0	0	397	0	0	0
displacement	mpg	Cor b/w mpg	397	0	0	397	0	0	0
mpg	mpg	Cor b/w mpg	397	0	0	397	0	0	0
weight	mpg	Cor b/w mpg	397	0	0	397	0	0	0
year	mpg	Cor b/w mpg	397	0	0	397	0	0	0

5 rows | 1-10 of 36 columns

3. Distributions of numerical variables

Graphical representation of all numeric features, used below types of plots to explore the data

- Quantile-quantile plot (Univariate)
- Density plot (Univariate)
- Scatter plot (Bivariate)

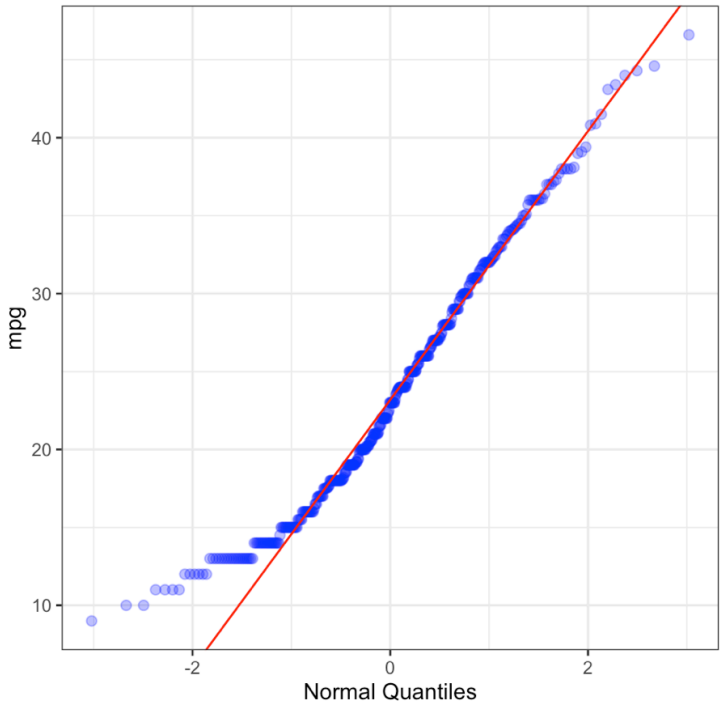
Quantile-quantile plot for Numerical variables - Univariate

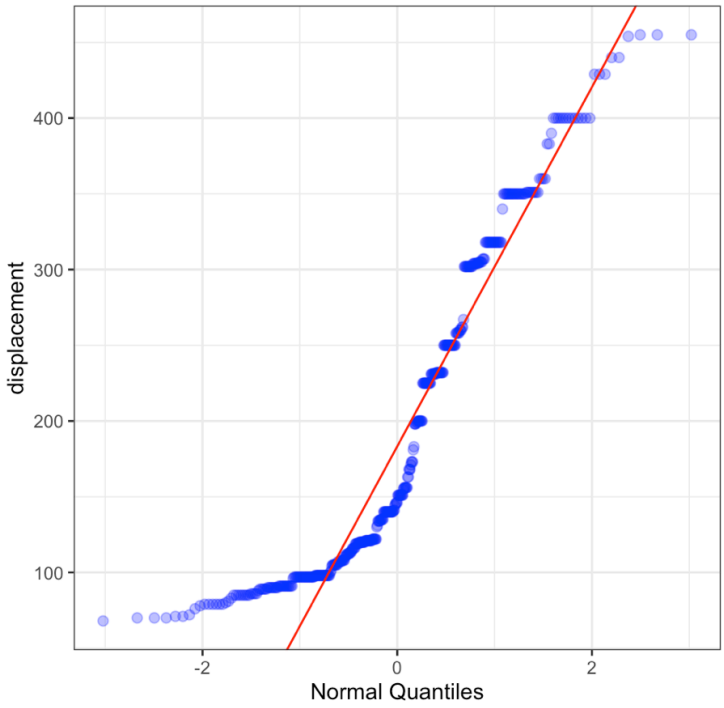
Quantile-quantile plot for all Numerical variables

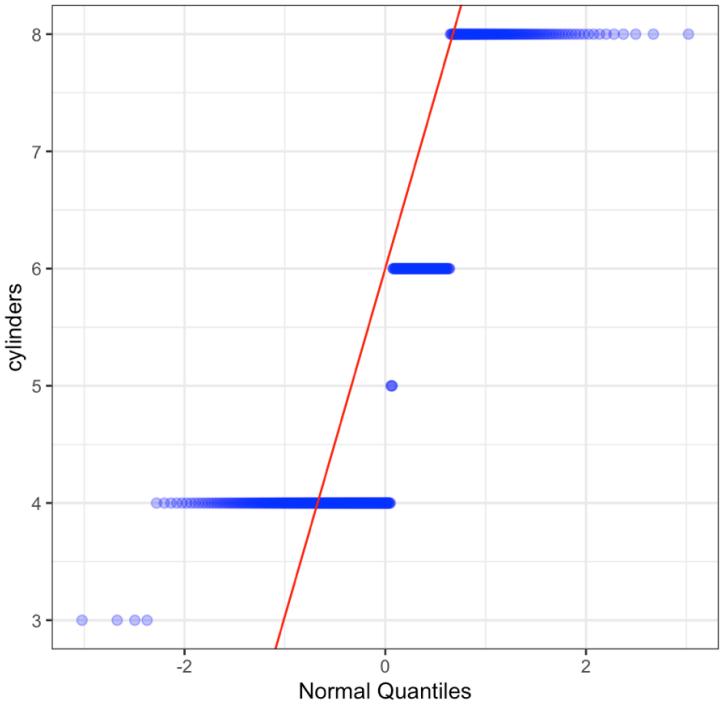
```
ExpOutQQ(data,nlim=4,fname=NULL,Page=c(2,2),sample=sn)
```

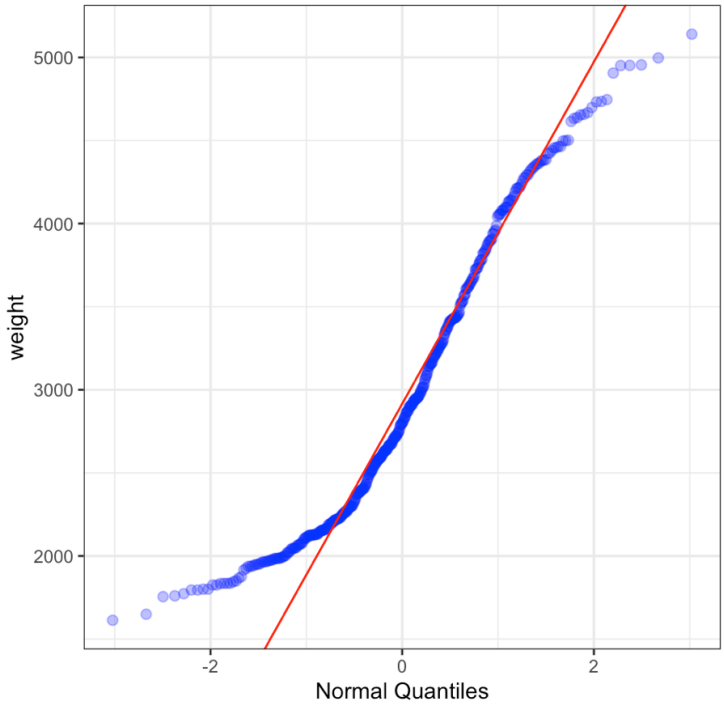
```
## $`0`
```

page 1 of 2

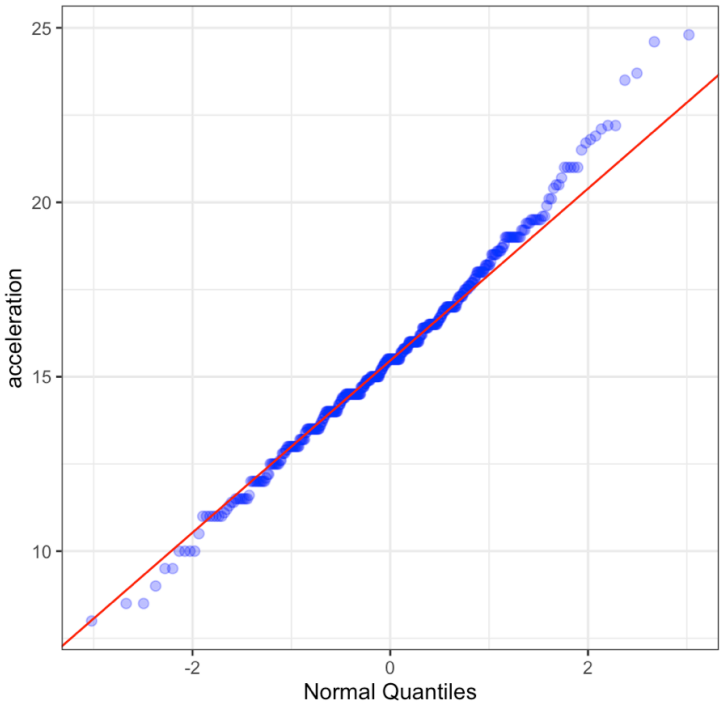

mpg
40
30
20
10
-2
0
2
Normal Quantiles

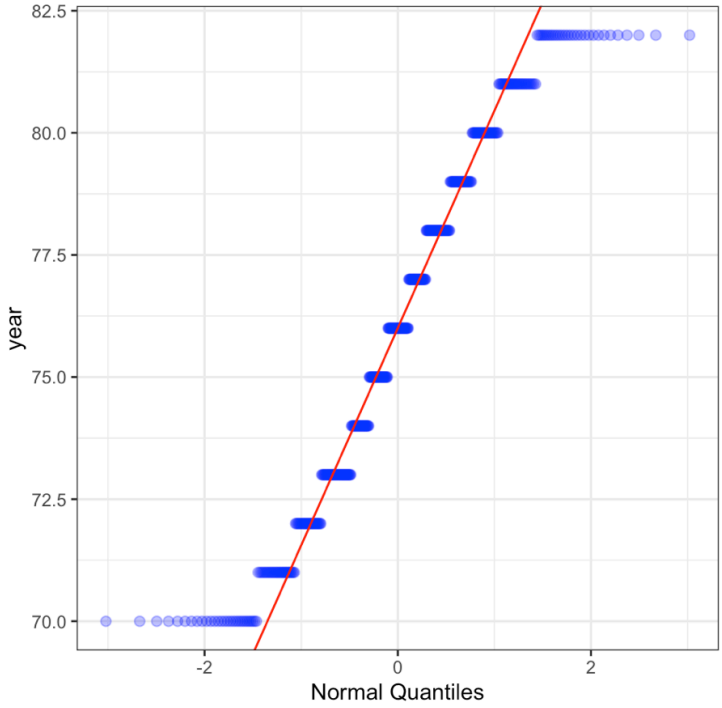

displacement
400
300
200
100
-2
0
2
Normal Quantiles


cylinders
8
7
6
5
4
3
-2
0
2
Normal Quantiles


weight
5000
4000
3000
2000
-2
0
2
Normal Quantiles

page 2 of 2





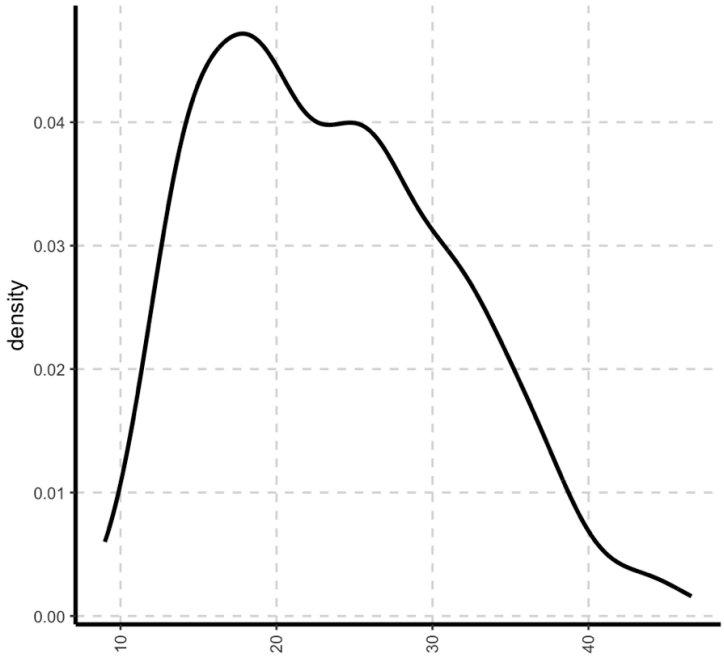
Density plots for numerical variables - Univariate

Density plot for all numerical variables

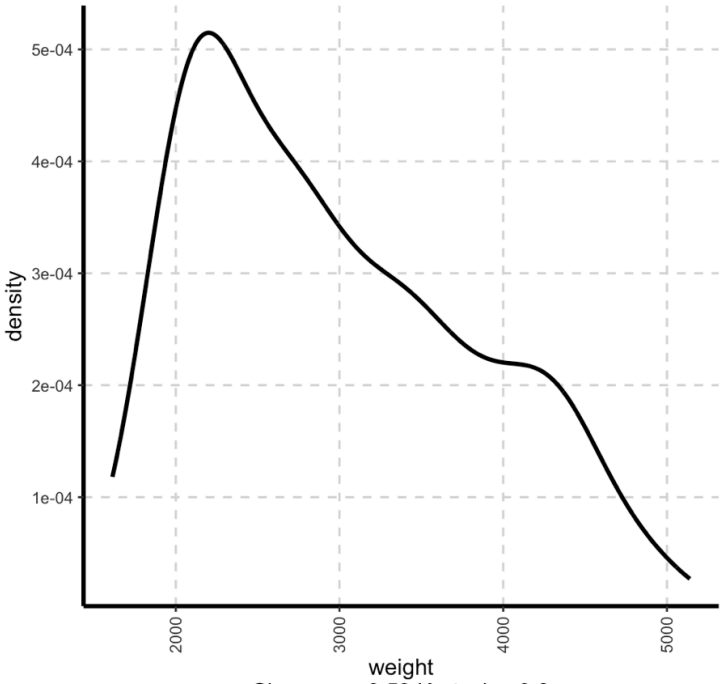
```
ExpNumViz(data,target=NULL,nlim=10,fname=NULL,col=NULL,theme=theme,Page=c(2,2),sample
=sn)
```

$`0`

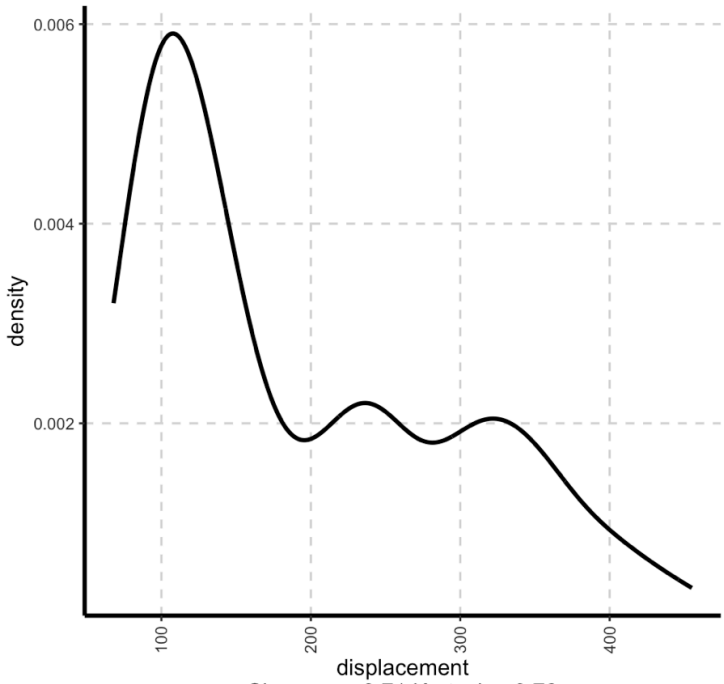
page 1 of 2



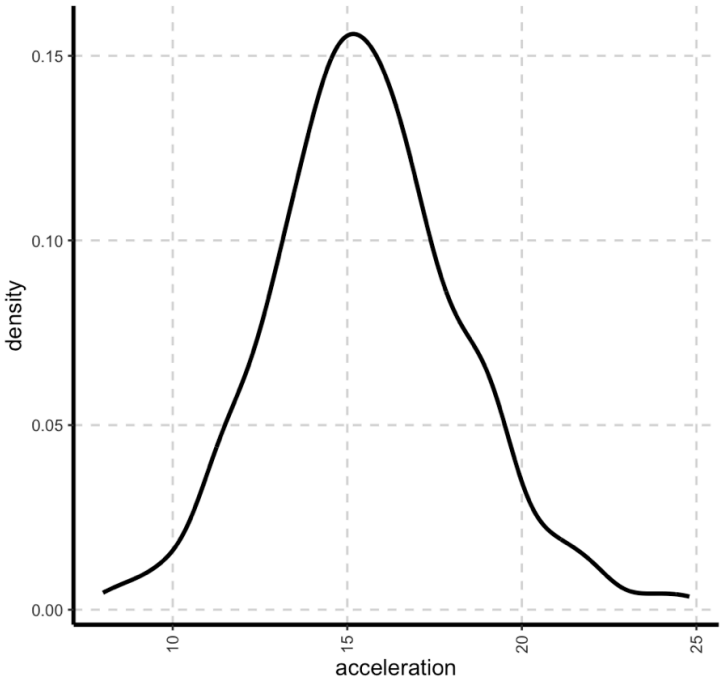
Skewness: 0.45 Kurtosis: -0.53



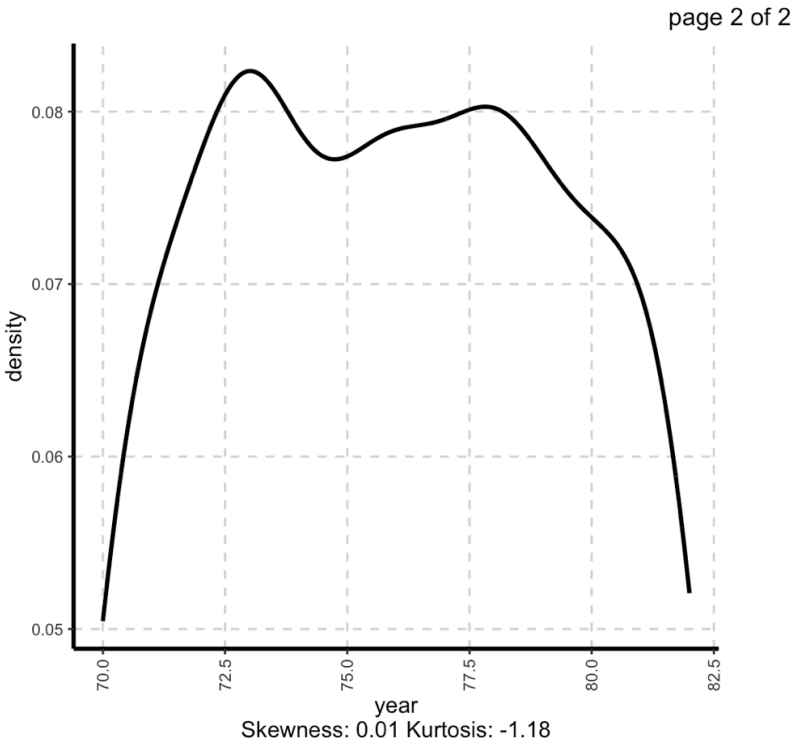
Skewness: 0.53 Kurtosis: -0.8



Skewness: 0.71 Kurtosis: -0.76



Skewness: 0.28 Kurtosis: 0.42

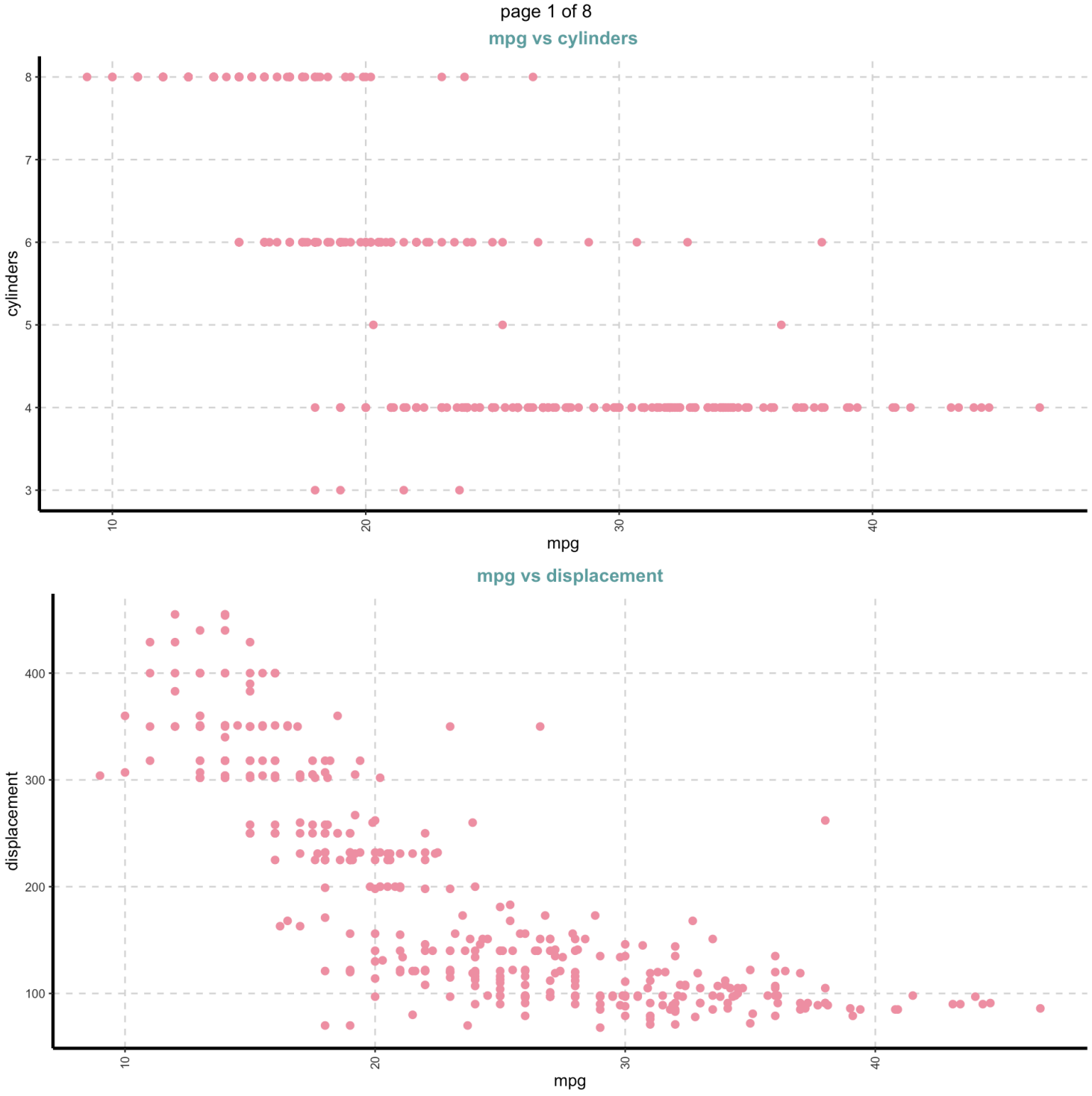


Scatter plot for all Numeric variables

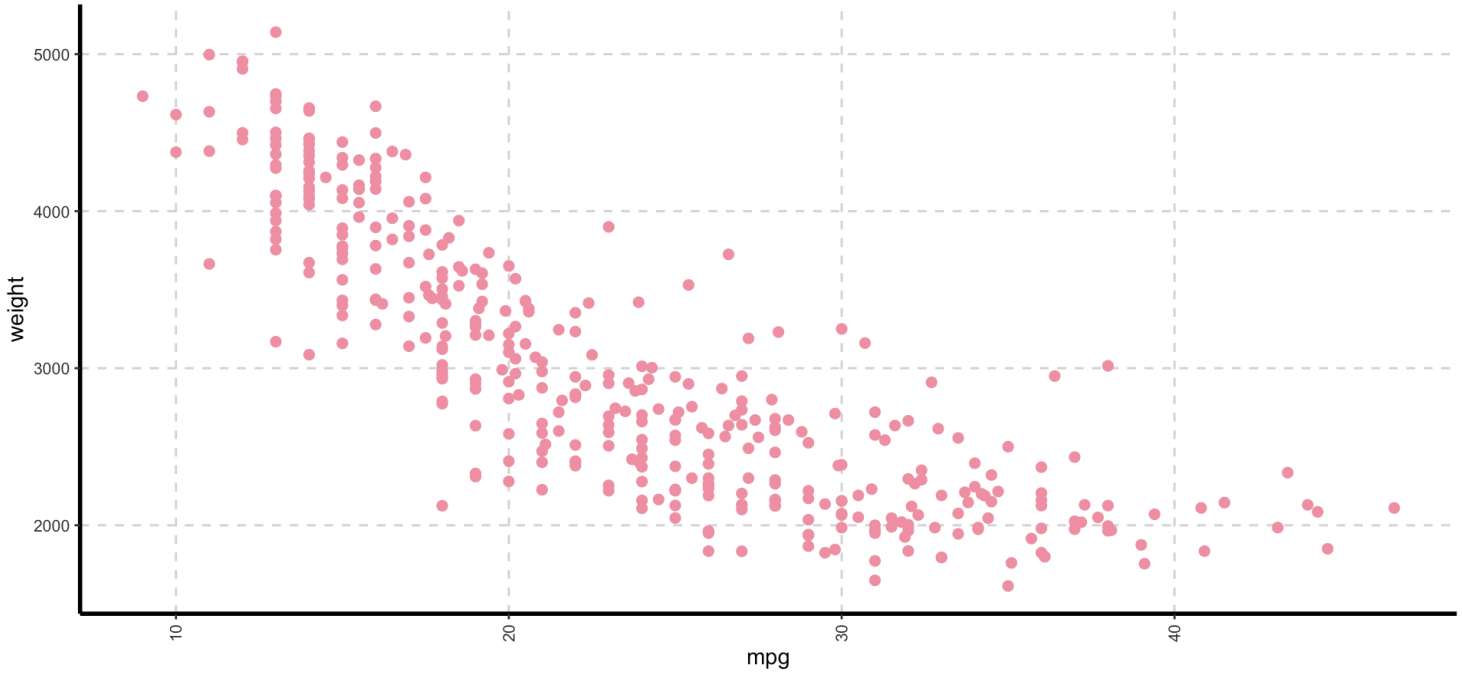
Scatter plot between all numeric variables and target variable **mpg**. This plot help to examine how well a target variable is correlated with list of dependent variables in the data set.

```
ExpNumViz(data,target=NULL,nlim=5,Page=c(2,1),theme=theme,sample=sn,scatter=TRUE)
```

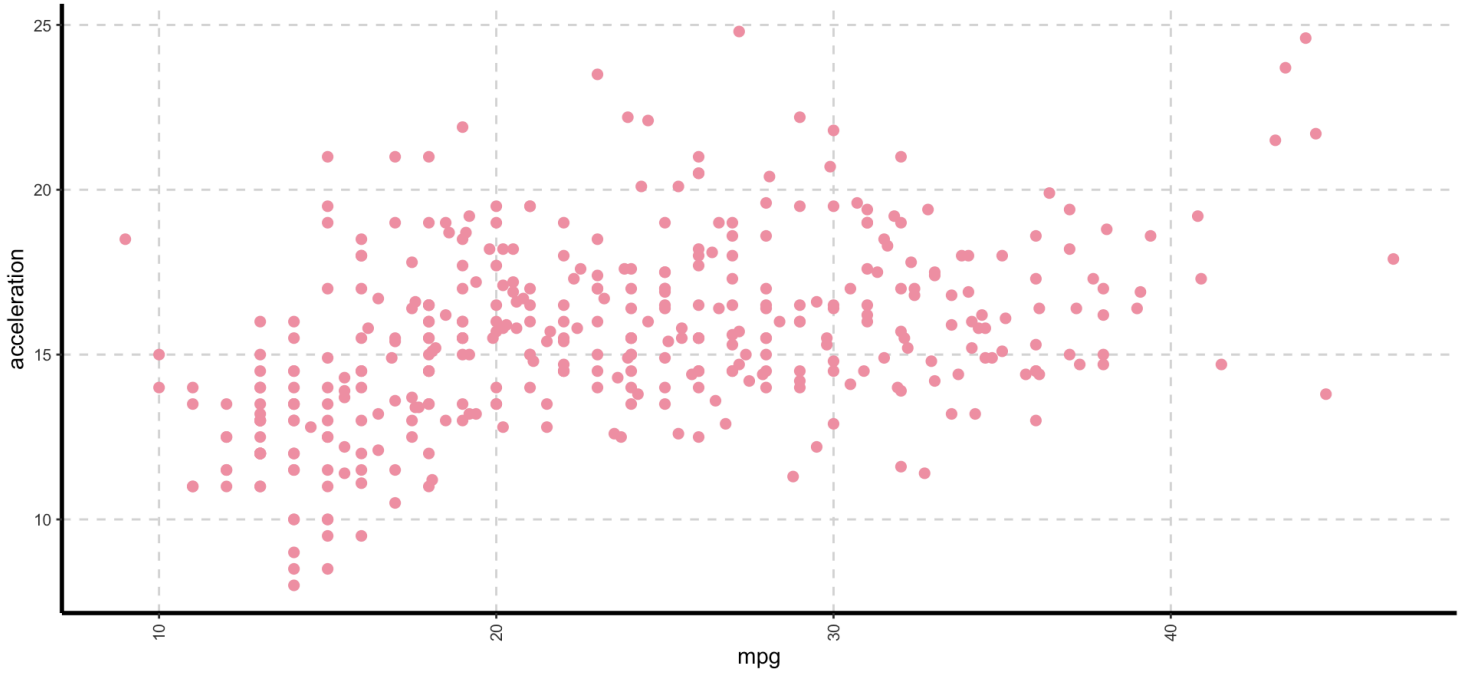
```
## $`0`
```


page 1 of 8
mpg vs cylinders
cylinders
8
7
6
5
4
3
10
20
30
40
mpg
mpg vs displacement
displacement
400
300
200
100
10
20
30
40
mpg

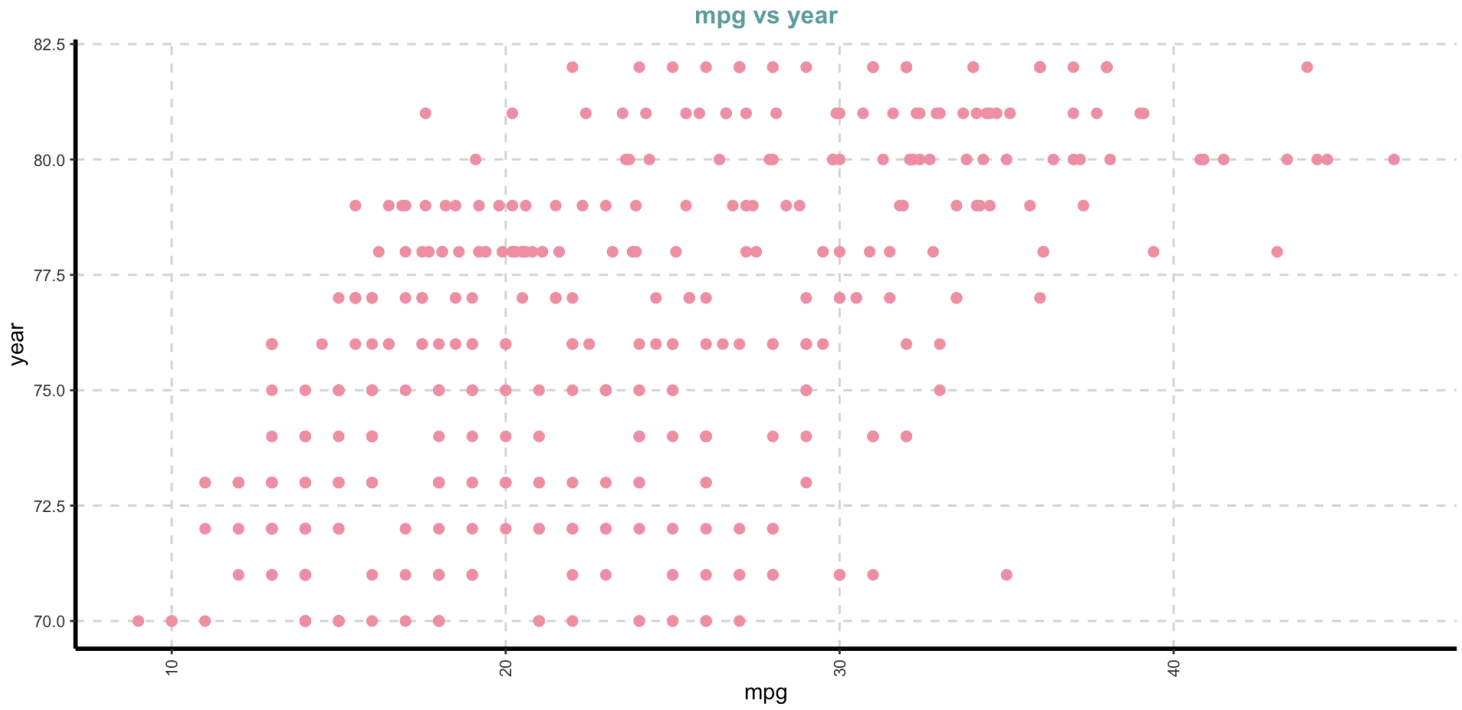
mpg vs weight

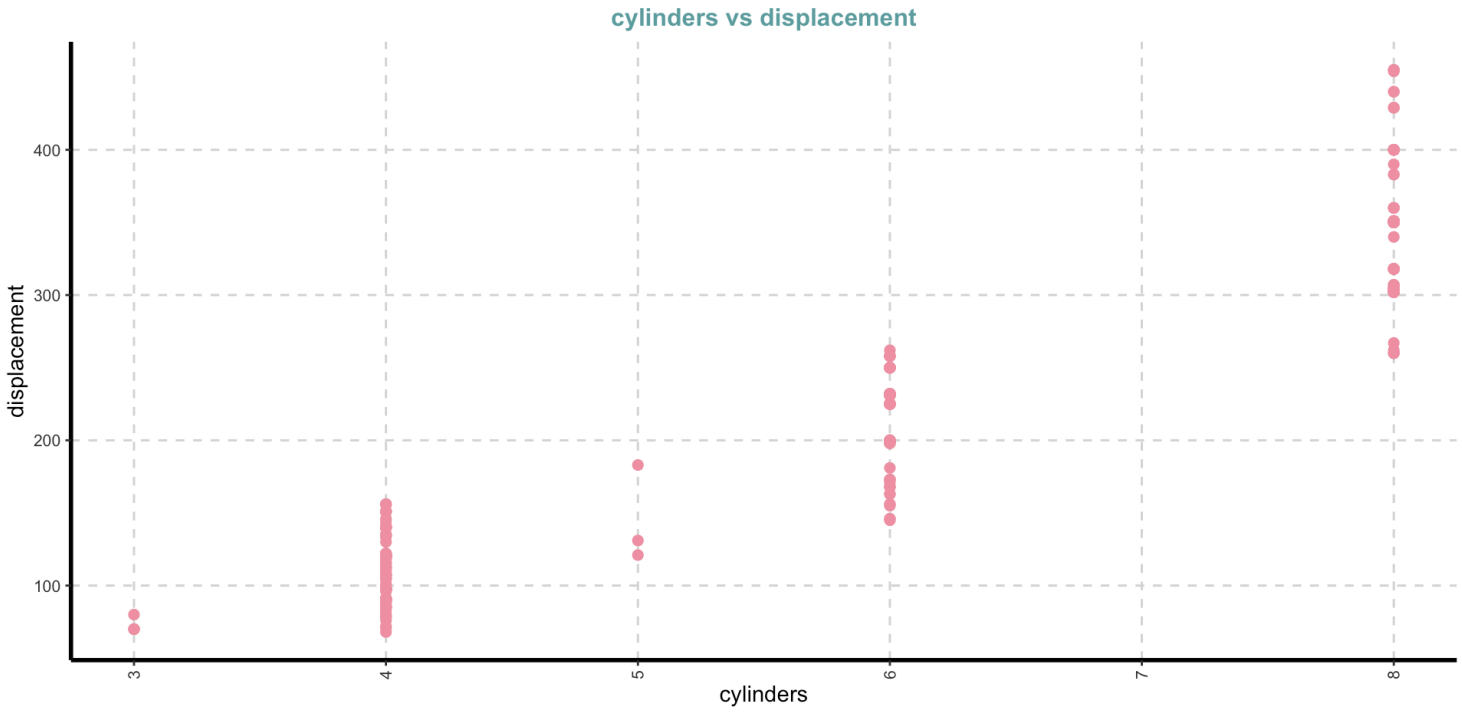


mpg vs acceleration



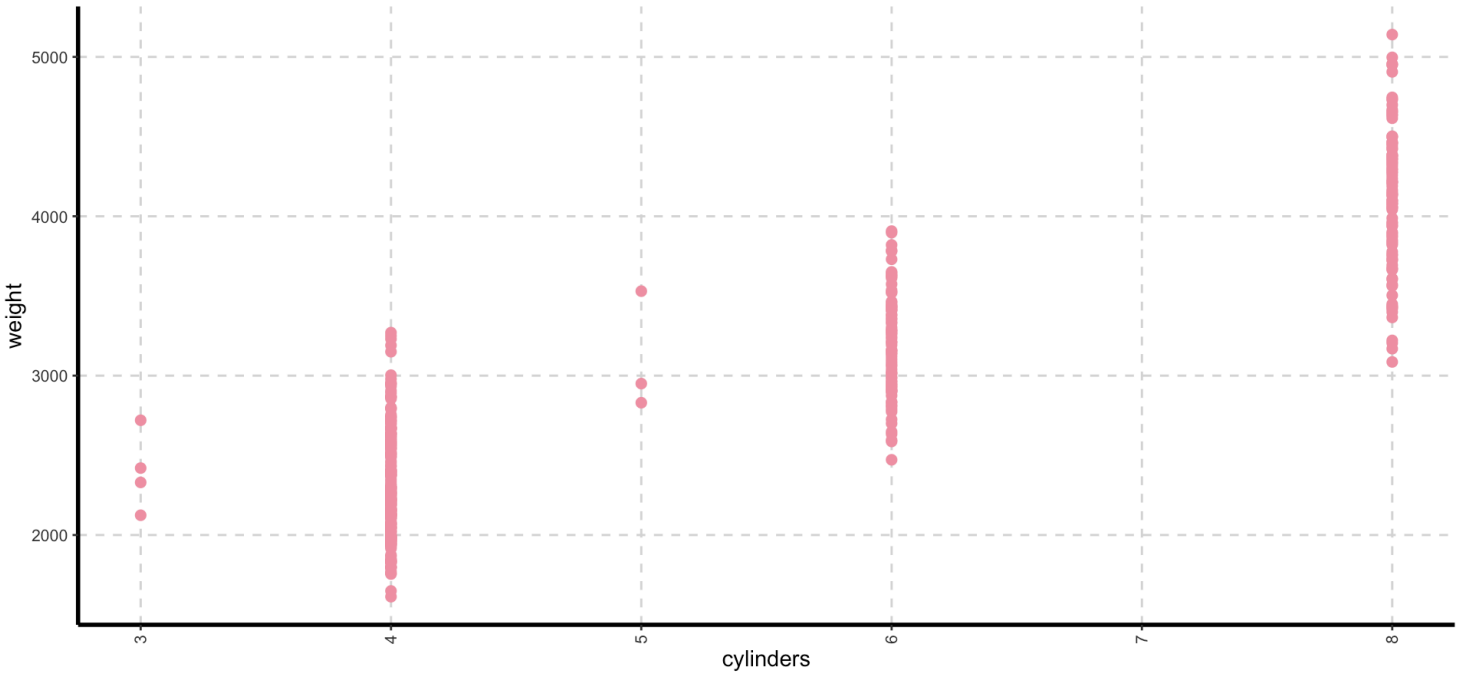
page 3 of 8


mpg vs year
year
mpg
82.5
80.0
77.5
75.0
72.5
70.0
10
20
30
40

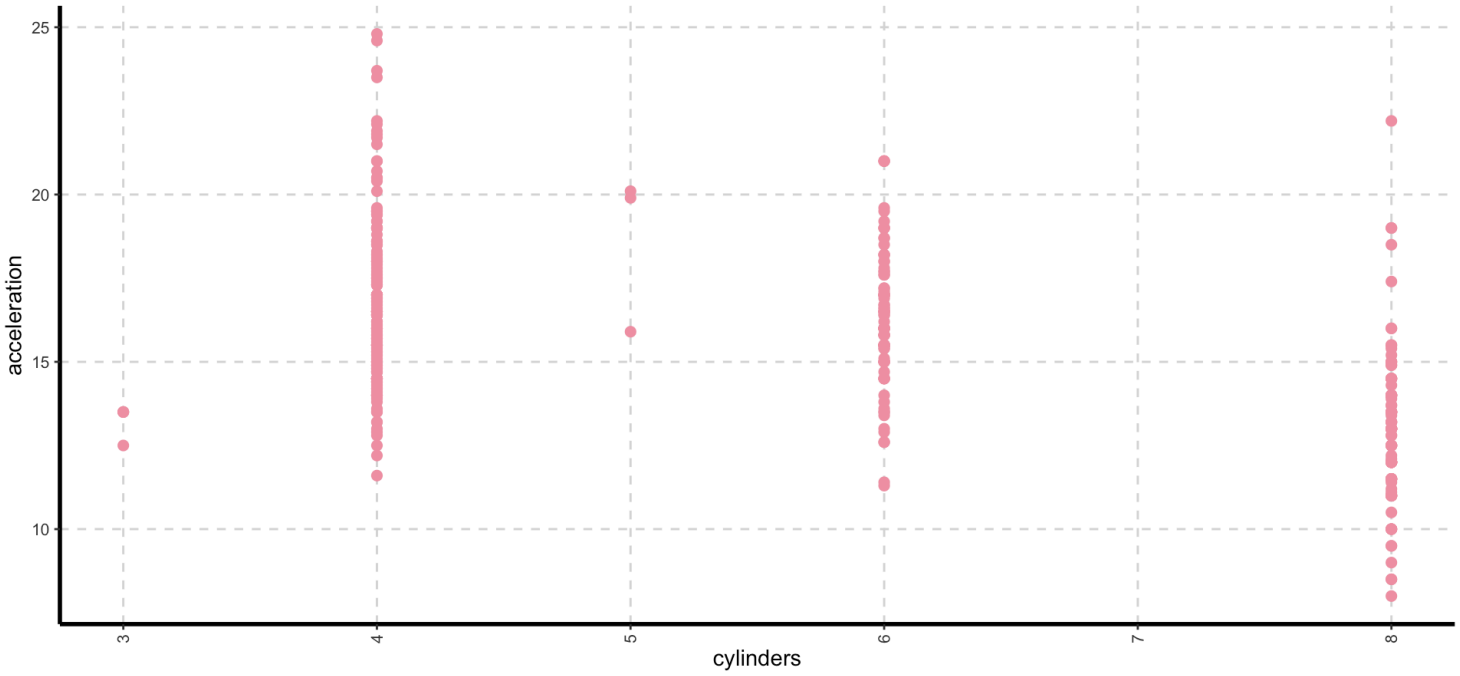

cylinders vs displacement
displacement
cylinders
400
300
200
100
3
4
5
6
7
8

page 4 of 8

cylinders vs weight

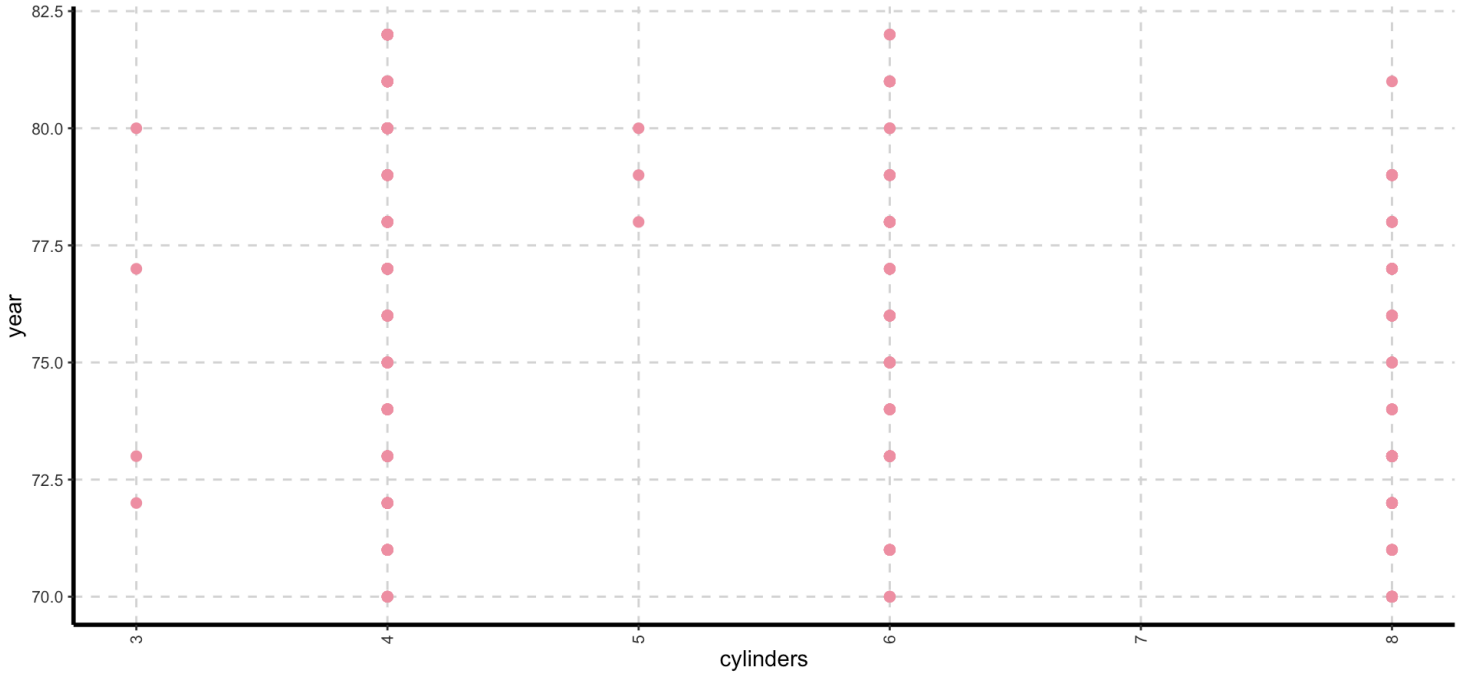


cylinders vs acceleration

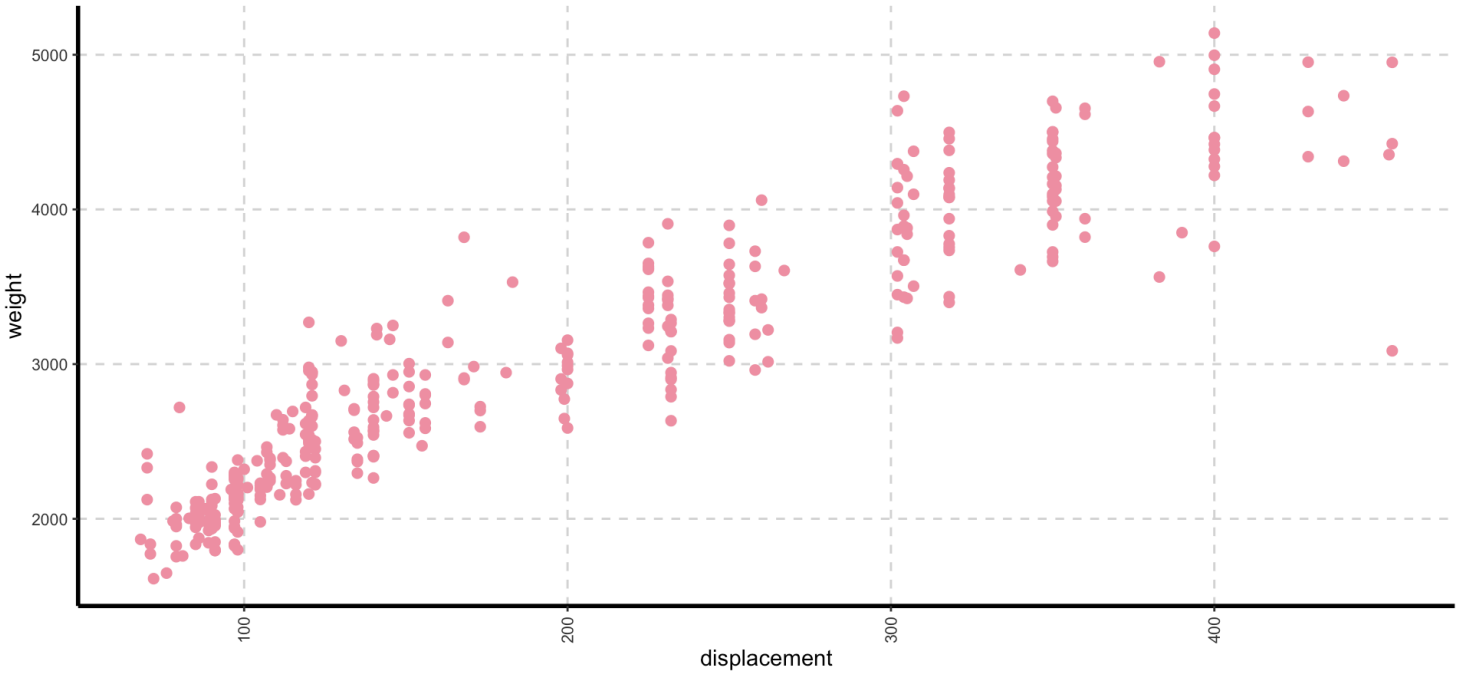


page 5 of 8

cylinders vs year



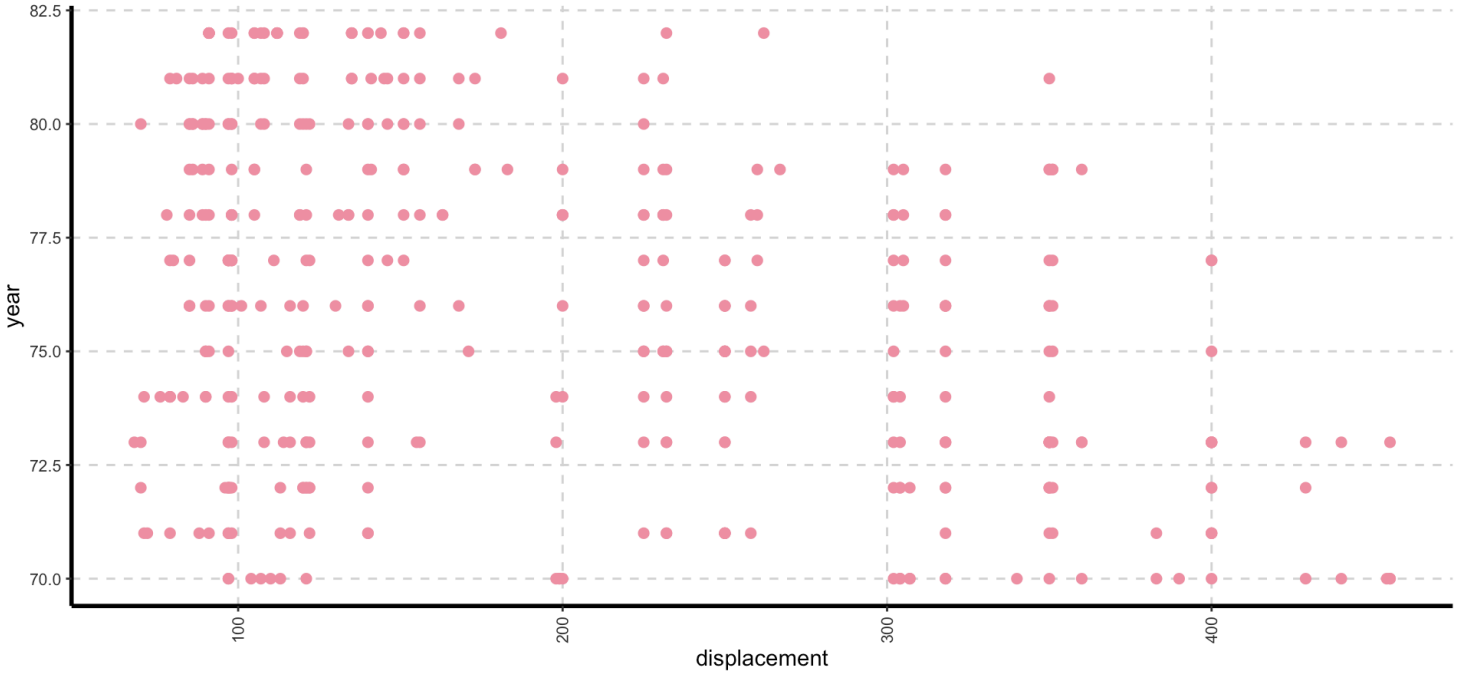
displacement vs weight



page 6 of 8

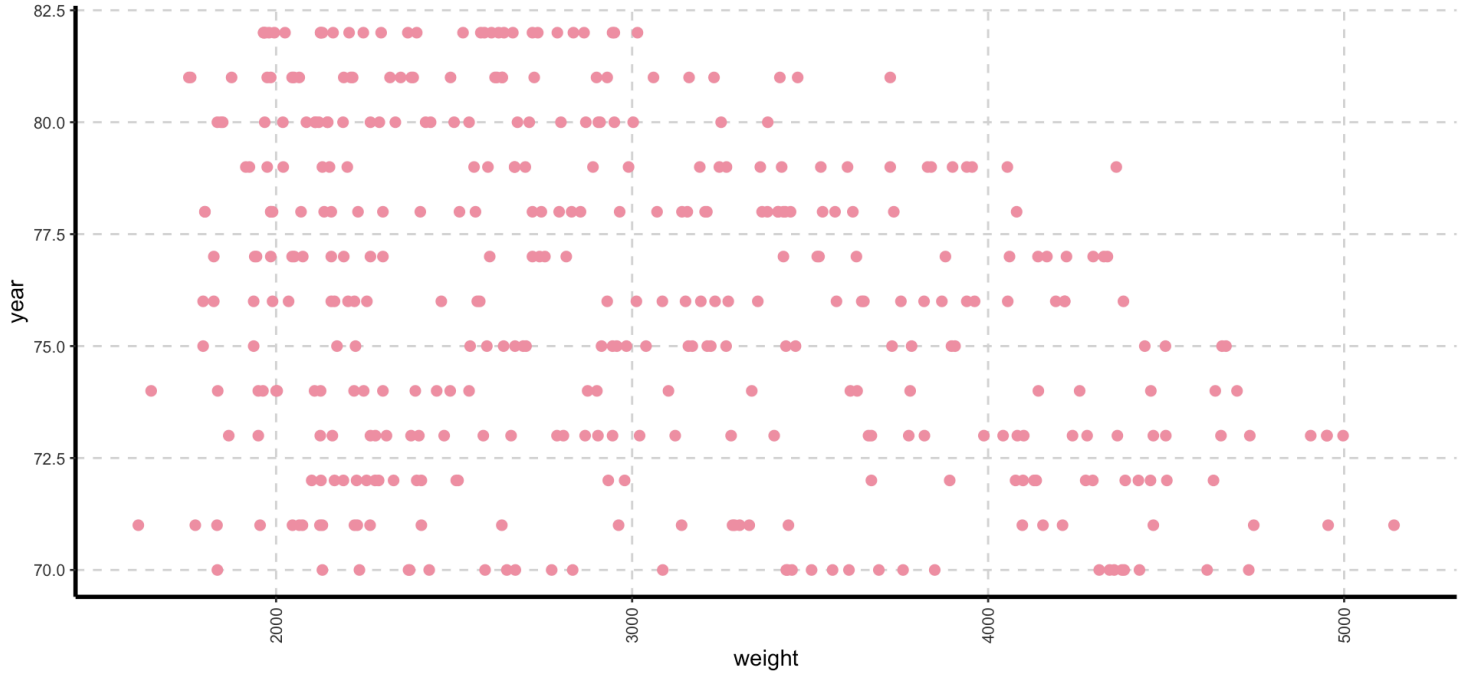
displacement vs acceleration

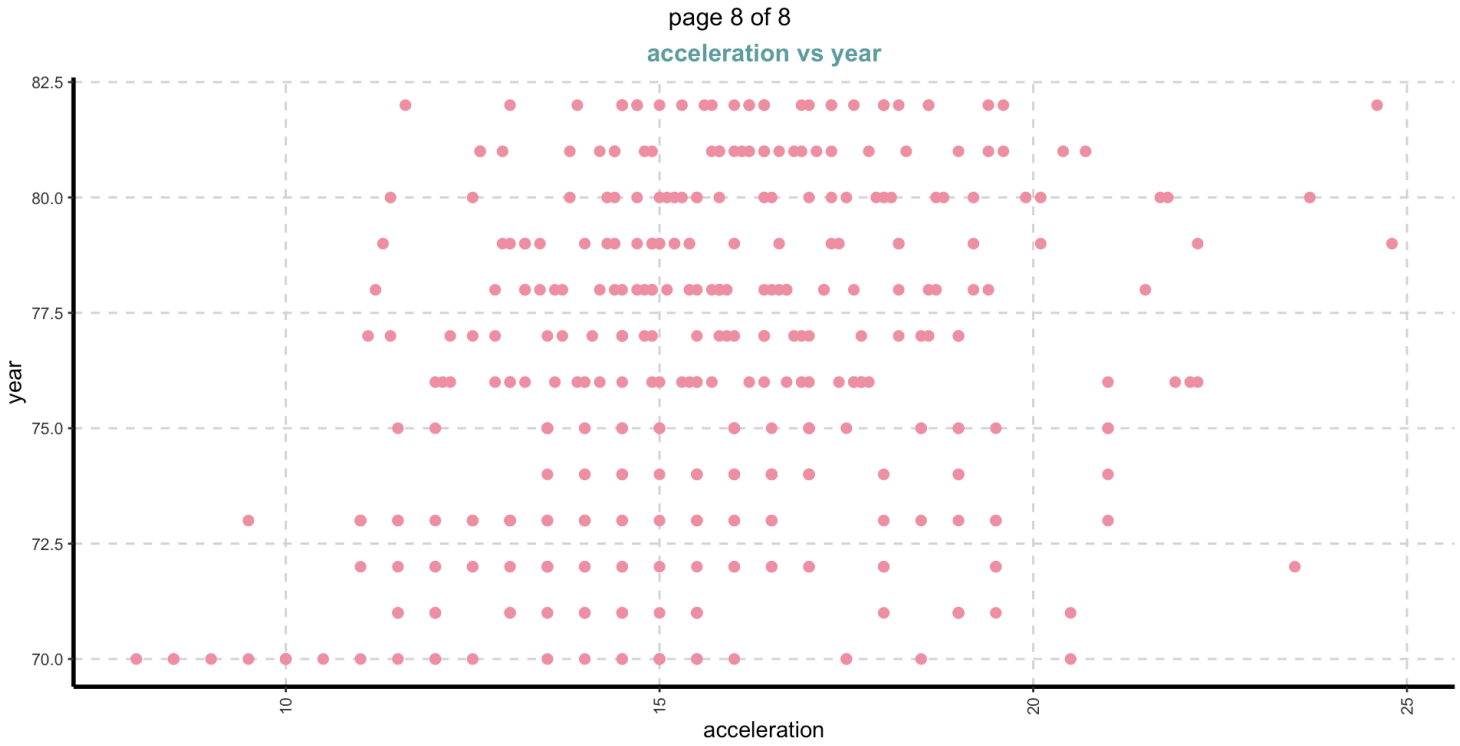
displacement vs year



weight vs acceleration

weight vs year



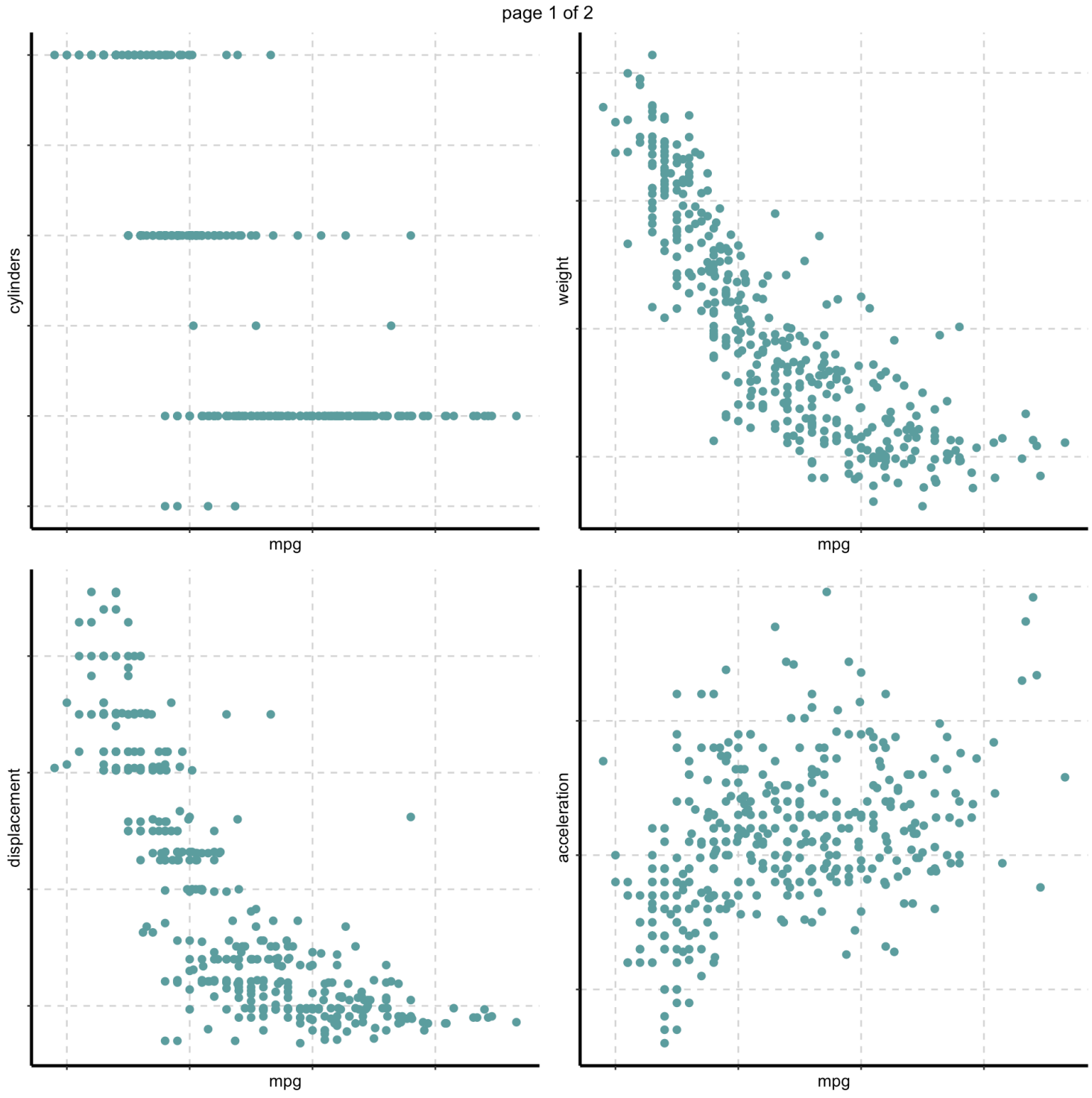


Correlation between dependent variable vs Independent variables

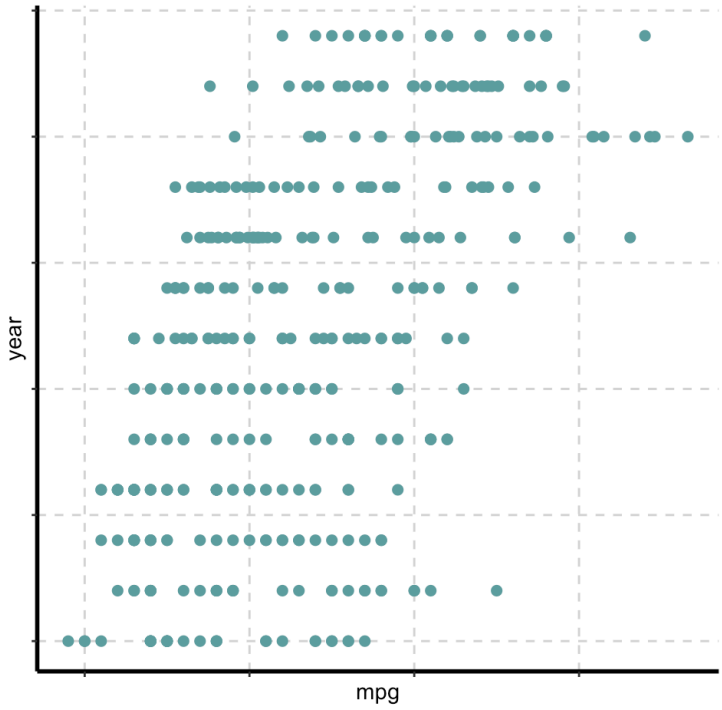
Dependent variable is **mpg** (continuous).

```
ExpNumViz(data,target=Target,nlim=5,fname=NULL,col=NULL,theme=theme,Page=c(2,2),sampl
e=sn)
```

$`0`


page 1 of 2
cylinders
mpg
weight
mpg
displacement
mpg
acceleration
mpg

page 2 of 2



** Correlation summary table

```
ExpNumStat(data,by="GA",gp=Target,MesofShape=2,Outlier=FALSE,round=2,dcast=T,val="cor
")
```

Stat	Vname	mpg
<chr>	<chr>	<dbl>

cor	acceleration	0.42
cor	displacement	-0.80
cor	mpg	1.00
cor	weight	-0.83
cor	year	0.58

5 rows

4. Summary of categorical variables

Summary of categorical variables

- frequency for all categorical independent variables

```
ExpCTable(data,margin=1,clim=10,nlim=5,round=2,per=T)
```

Variable <chr>	**Valid** <chr>	**Frequency** <dbl>	**Percent** <dbl>	**CumPercent** <dbl>
cylinders	3	4	1.01	1.01
cylinders	4	203	51.13	52.14
cylinders	5	3	0.76	52.90
cylinders	6	84	21.16	74.06
cylinders	8	103	25.94	100.00
cylinders	TOTAL	397	*NA*	*NA*
origin	1	248	62.47	62.47
origin	2	70	17.63	80.10
origin	3	79	19.90	100.00
origin	TOTAL	397	*NA*	*NA*

1-10 of 10 rows

- frequency for all categorical independent variables by descretized **mpg**

```
##bin=4, descretized 4 categories based on quantiles
ExpCTable(data,Target=Target,margin=1,clim=10,nlim=5,round=2,bin=4,per=T)
```

VARIABLE <chr>	CATEG... <chr>	Nu... <chr>	mpg:(8.96,18.4] <dbl>	mpg:(18.4,27.8] <dbl>	mpg:(27.8,37.2] <dbl>	mpg:(
cylinders	3	nn	1.00	3.00	0	
cylinders	4	nn	1.00	88.00	96	
cylinders	5	nn	0.00	2.00	1	
cylinders	6	nn	32.00	48.00	3	
cylinders	8	nn	93.00	10.00	0	
cylinders	TOTAL	nn	127.00	151.00	100	
cylinders	3	%	0.79	1.99	0	
cylinders	4	%	0.79	58.28	96	
cylinders	5	%	0.00	1.32	1	
cylinders	6	%	25.20	31.79	3	

1-10 of 20 rows Previous **1** 2 Next

5. Distributions of Categorical variables

Graphical representation of all Categorical variables

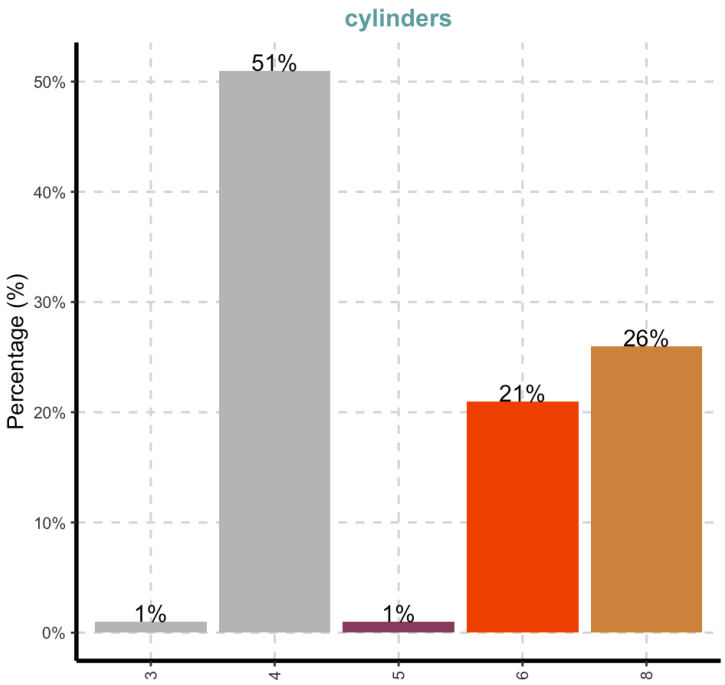
- Bar plot (Univariate)

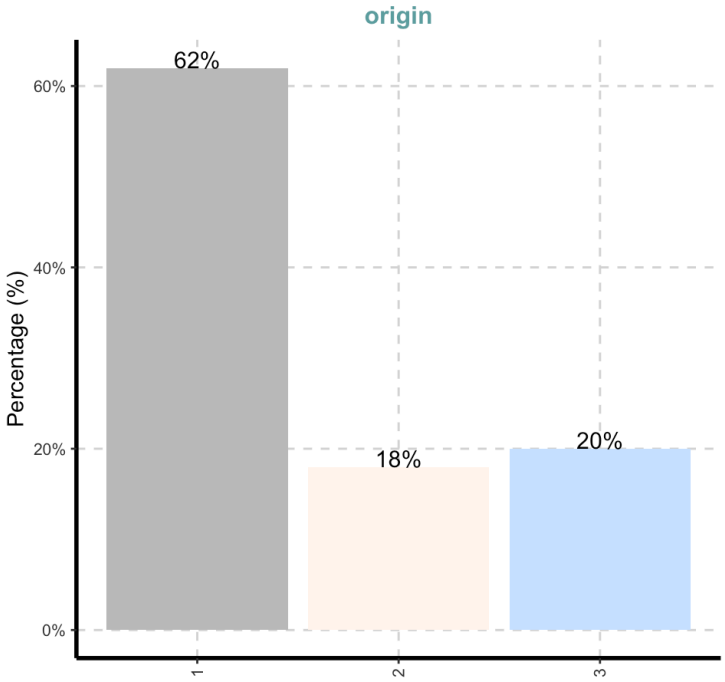
Bar plot with vertical or horizontal bars for all categorical variables

```
ExpCatViz(data,clim=10,margin=2,theme=theme,Page = c(2,2),sample=sc)
```

```
## $`0`
```

page 1 of 1


cylinders
Percentage (%)
50%
40%
30%
20%
10%
0%
51%
26%
21%
1%
1%
3
4
5
6
8


origin
Percentage (%)
60%
40%
20%
0%
62%
18%
20%
1
2
3