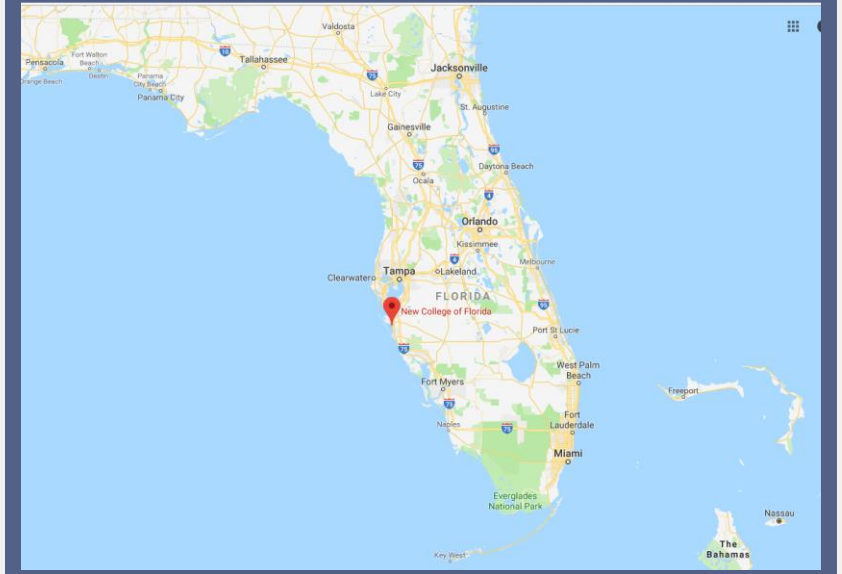
HANDLING MISSING DATA IN R

Mine Dogucu, PhD
Sunwoo Ha
New College of Florida

@MineDogucu
mdogucu
mdogucu@ncf.edu

New College
THE HONORS COLLEGE of Florida



	Fall	Spring
Year 1	Statistical Inference I	Statistical Inference II
	Data Storage and Retrieval	Data Visualization
	Algorithms	Distributed Computing
	Data Munging and EDA	Optimization and Machine Learning
Year 2	Practical Data Science	Practicum
	Topics in Computing - Deep Learning	
	Topics in Statistical Inference - Applied Bayesian Analysis	

	Fall	Spring
Year 1	Statistical Inference I	Statistical Inference II
	Data Storage and Retrieval	Data Visualization
	Algorithms	Distributed Computing
	Data Munging and EDA	Optimization and Machine Learning
Year 2	Practical Data Science	Practicum
	Topics in Computing - Deep Learning	
	Topics in Statistical Inference - Applied Bayesian Analysis	

```
> library(ggplot2movies)
> View(movies)
```

	title	year	length	budget	rating	votes	r1	r2
272	20-seiki nosutarujia	1997	93	NA	3.7	14	24.5	0
273	20. Juli, Der	1955	97	NA	8.6	18	4.5	0
274	20/20 Vision	1999	20	NA	1.0	5	64.5	0
275	200 American	2003	84	NA	5.4	62	14.5	4
276	200 Cigarettes	1999	101	6000000	5.4	4514	4.5	4
277	200 Motels	1971	98	679000	5.2	338	4.5	4
278	2000 Nordestes	2000	70	NA	7.9	25	4.5	0
279	2000 Years Later	1969	80	NA	4.4	12	4.5	14
280	20000 Leagues Under the Sea	1954	127	5000000	7.1	2741	4.5	4
281	2001 Yonggary	1999	99	NA	3.1	241	44.5	14
282	2001: A Space Odyssey	1968	156	10500000	8.3	64982	4.5	4
283	2001: A Space Travesty	2000	99	26000000	2.5	2023	44.5	14
284	2002: The Rape of Eden	1994	90	NA	4.0	24	14.5	4
285	2009: Lost Memories	2002	136	NA	6.6	639	4.5	4
286	201 Kanarinia, Ta	1964	96	NA	7.1	6	0.0	0
287	2010	1984	116	NA	6.5	7300	4.5	4

```
> library(janeaustenr)
> View(prideprejudice)
```

1	PRIDE AND PREJUDICE
2	
3	By Jane Austen
4	
5	
6	
7	Chapter 1
8	
9	
10	It is a truth universally acknowledged, that a single man ...
11	of a good fortune, must be in want of a wife.
12	
13	However little known the feelings or views of such a ma...
14	first entering a neighbourhood, this truth is so well fixe...
15	of the surrounding families, that he is considered the ri...
16	of some one or other of their daughters.

"As the old saying goes, the only certainties are death and taxes. We would like to add one more to that list: missing data".

McKnight et. al (2007)

SELECTIVE NONRESPONSE

You love your everyday job which involves data analysis.

AGREE ○ ○ ○ ○ ○ DISAGREE

You hate doing data analysis when it involves missing data

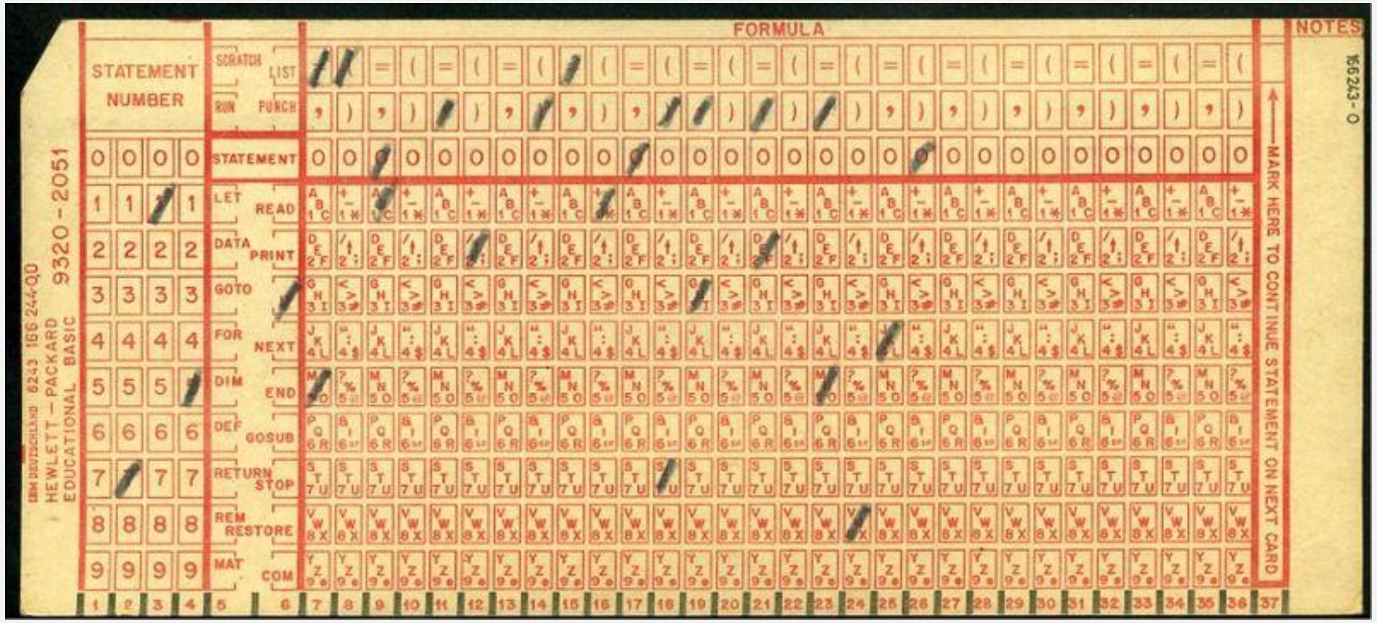
AGREE ○ ○ ○ ○ ○ DISAGREE

DROPOUT

1 2 3 4 5

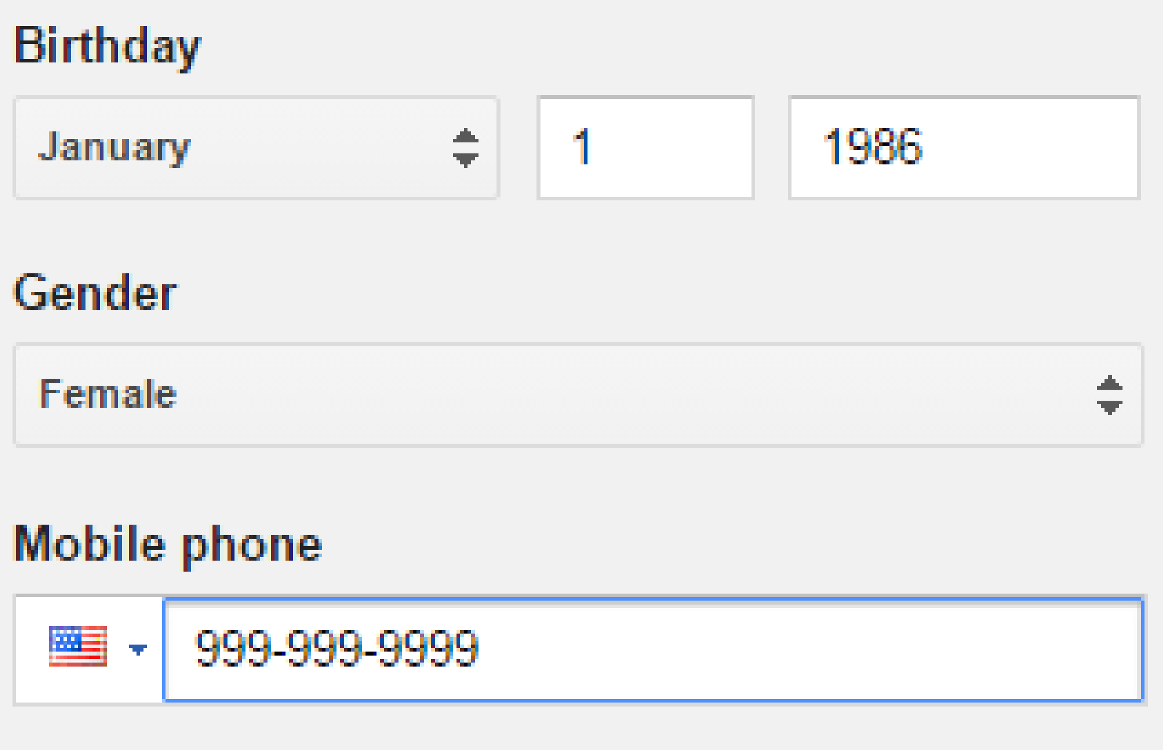
WAVE MISSING

TECHNOLOGY



https://upload.wikimedia.org/wikipedia/commons/9/95/HP_Educational_Basic_optical_mark-reader_card._Godfrey_Manning..jpg

INVALID RESPONSES



PLANNED MISSING DESIGNS

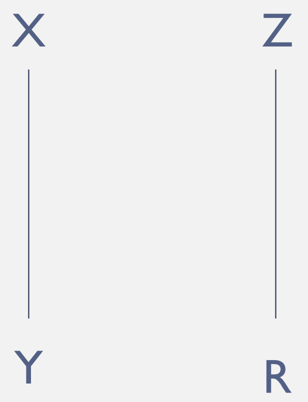
MISSING DATA MECHANISMS

- X = completely observed variable (s)
- Y = partly observed variable (s)
- Z = unobserved variables (unrelated to X and Y)
- R = indicates missingness

Rubin (1987)

MCAR

Students forget to fill out the survey

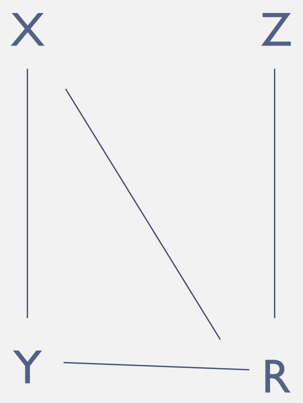


MAR

Younger students have hard time reading / understanding the questions

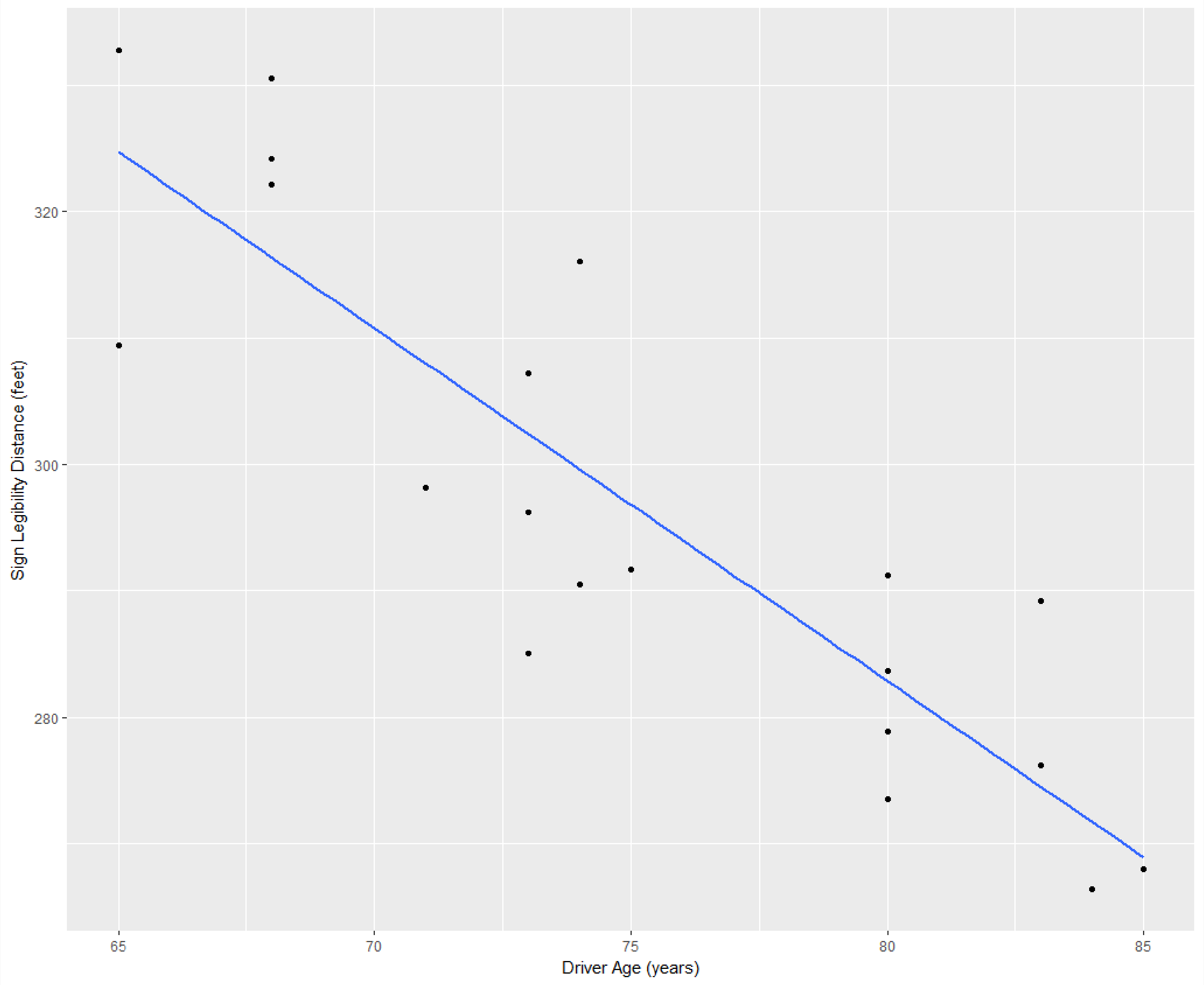
MNAR

Students with severe asthma do not have the energy to respond.



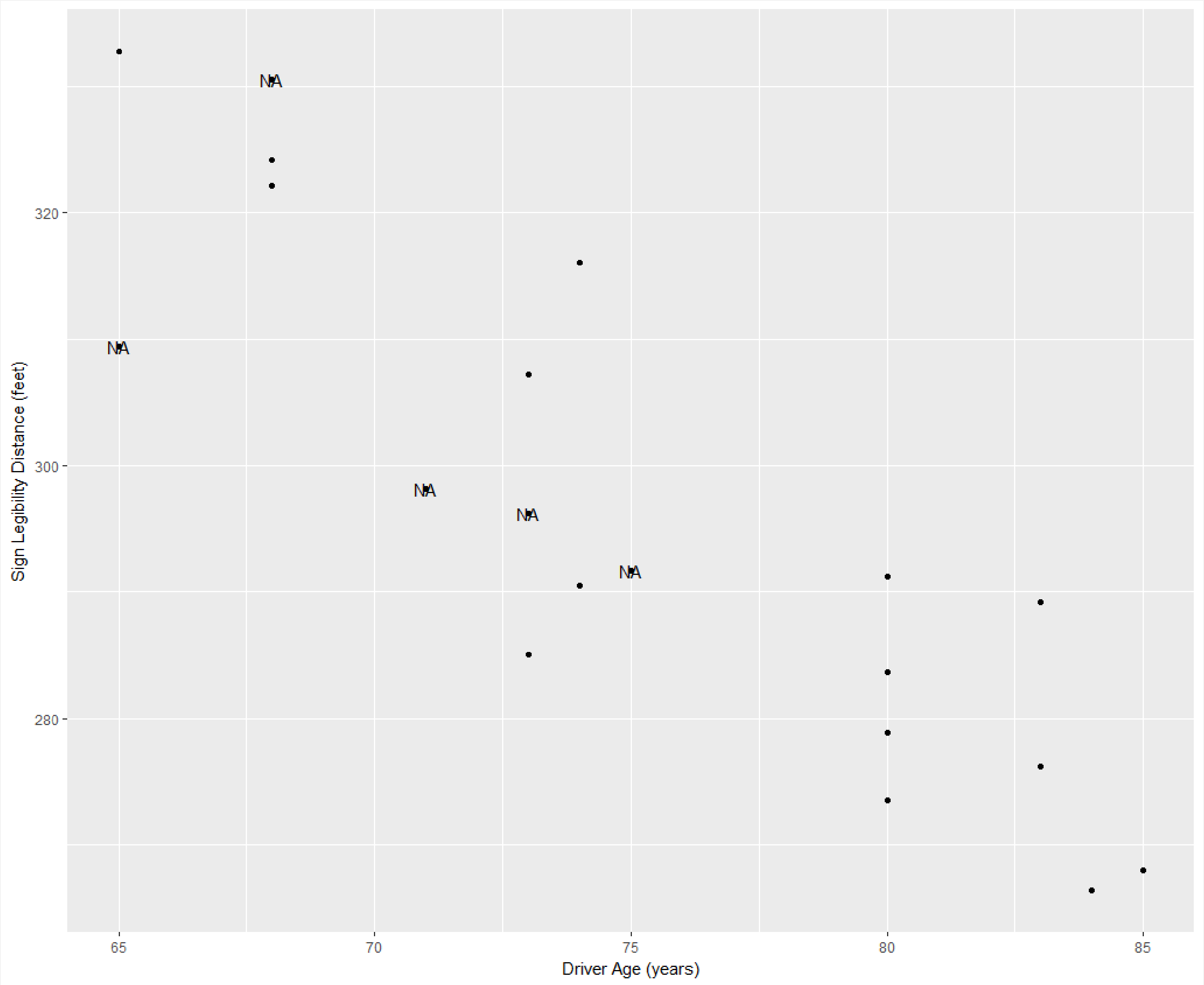
Schafer & Graham (2002)

$$Y_{Distance}=\beta_0 + \beta_1(Age) + \mathcal{E}$$



	age	distance
1	80	278.8375
2	83	289.1731
3	80	283.7063
4	83	276.2022
5	74	290.4947
6	68	324.1690
7	71	298.1364
8	75	291.6842
9	80	291.2071
10	85	267.9872
11	65	332.7962
12	68	330.5579
13	80	273.5567
14	65	309.4686
15	73	285.0229
16	74	316.0510
17	73	296.1835
18	73	307.2038
19	68	322.1212
20	84	266.3769

MAR



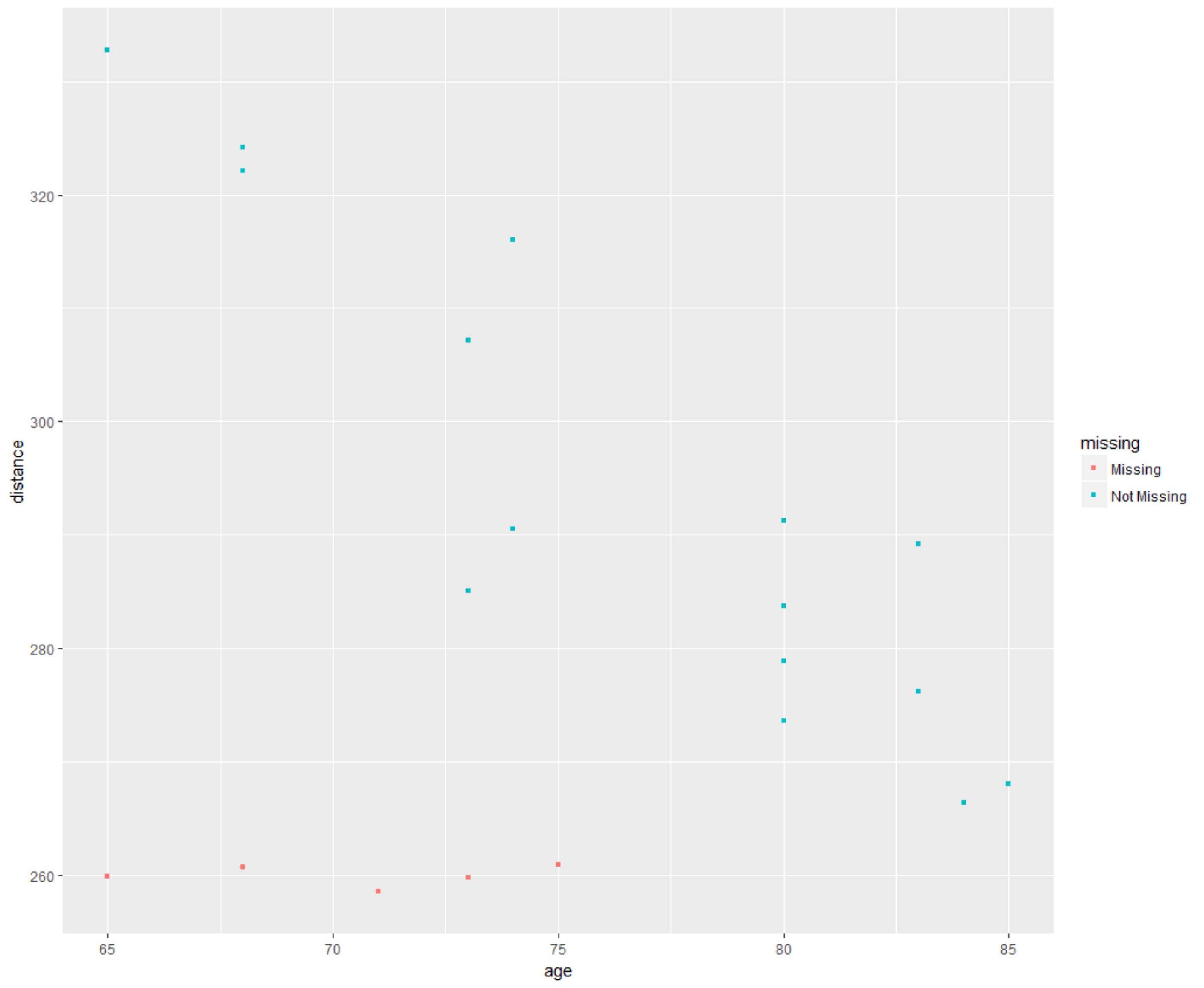
	age	distance	r
1	80	278.8375	
2	83	289.1731	
3	80	283.7063	
4	83	276.2022	
5	74	290.4947	
6	68	324.1690	
7	71	NA	NA
8	75	NA	NA
9	80	291.2071	
10	85	267.9872	
11	65	332.7962	
12	68	NA	NA
13	80	273.5567	
14	65	NA	NA
15	73	285.0229	
16	74	316.0510	
17	73	NA	NA
18	73	307.2038	
19	68	322.1212	
20	84	266.3769	

bit.ly/MissingDataR

SUMMARIES OF MISSING DATA

Initial step in any data analysis with missing data analysis should include visual and numerical inspection.

library(naniar)




```
ggplot(data = miss,
       aes(x = age,
           y = distance)) +
  geom_miss_point()
```

```
> miss_case_summary(miss)
# A tibble: 20 x 4
    case n_miss pct_miss n_miss_cumsum
 * <int>  <int>    <dbl>         <int>
 1     1      0      0               0
 2     2      0      0               0
 3     3      0      0               0
 4     4      0      0               0
 5     5      0      0               0
 6     6      0      0               0
 7     7      1     50.0             1
 8     8      1     50.0             2
 9     9      0      0               2
10    10      0      0               2
11    11      0      0               2
12    12      1     50.0             3
13    13      0      0               3
14    14      1     50.0             4
15    15      0      0               4
16    16      0      0               4
17    17      1     50.0             5
18    18      0      0               5
19    19      0      0               5
20    20      0      0               5
```

LITTLE'S MCAR TEST

$$H_0: Data\ are\ MCAR$$

```
> library(BaylorEdPsych)
> library(mvnmle)
> LittleMCAR(miss)
this could take a while$chi.square
[1] 3.74819

$df
[1] 1

$p.value
[1] 0.05286473

$missing.patterns
[1] 2

$amount.missing
               age distance
Number Missing   0     5.00
Percent Missing  0     0.25
...
```

TRUE PARAMETERS

$$\beta_0 = 500$$
$$\beta_1 = -3$$

SAMPLE STATISTICS

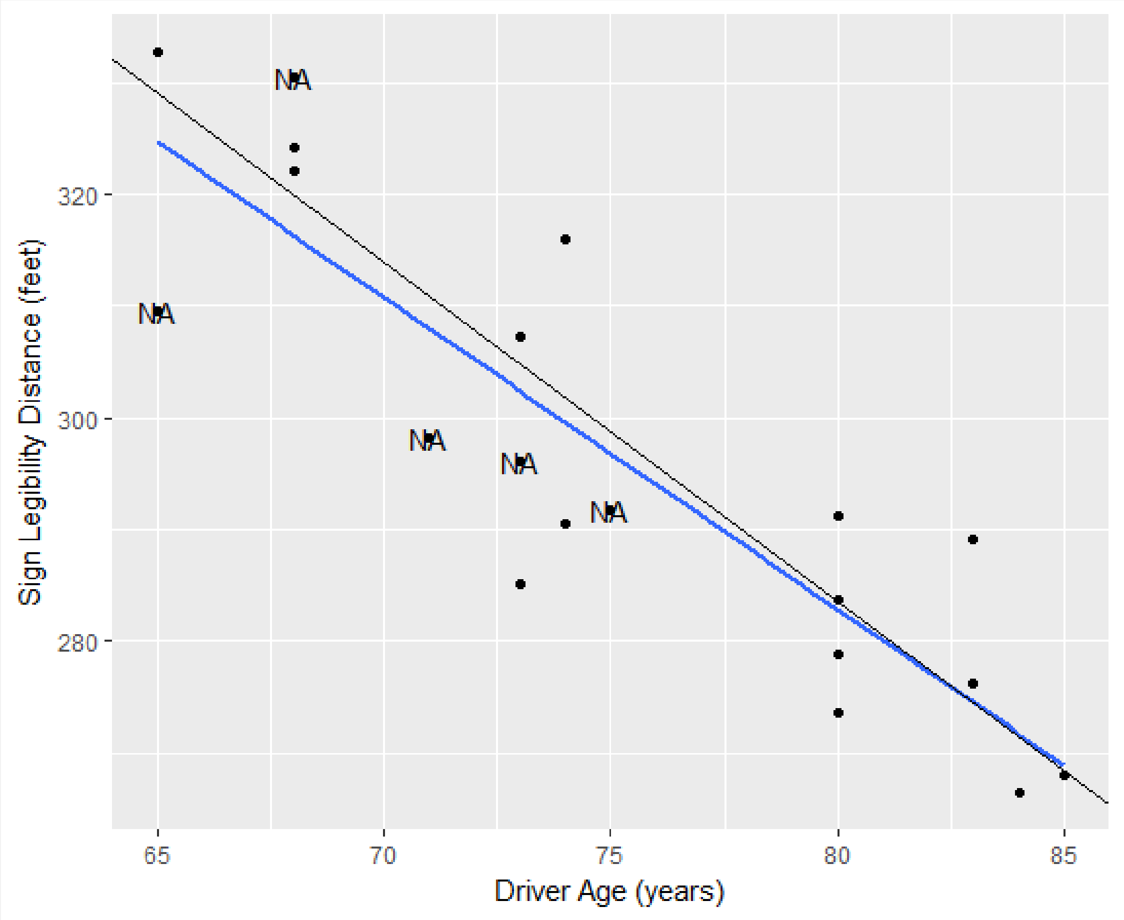
```
> coef(lm(distance~age,data=comp))
(Intercept)         age
 506.114401   -2.790514
```

COMPLETE CASE ANALYSIS

While analyzing, it includes only the rows that have complete data.

Default setting in many software (SPSS, SAS, STATA) but that is not always the case in R. e.g.

```
> mean(miss$distance)
[1] NA
> mean(miss$distance,na.rm = TRUE)
[1] 293.6604
```




```
> coef(lm(distance~age,data=miss))
(Intercept)         age
 526.551753   -3.037713
```

MEAN IMPUTATION

```
> mean(miss$distance,na.rm=TRUE)
[1] 293.6604
```

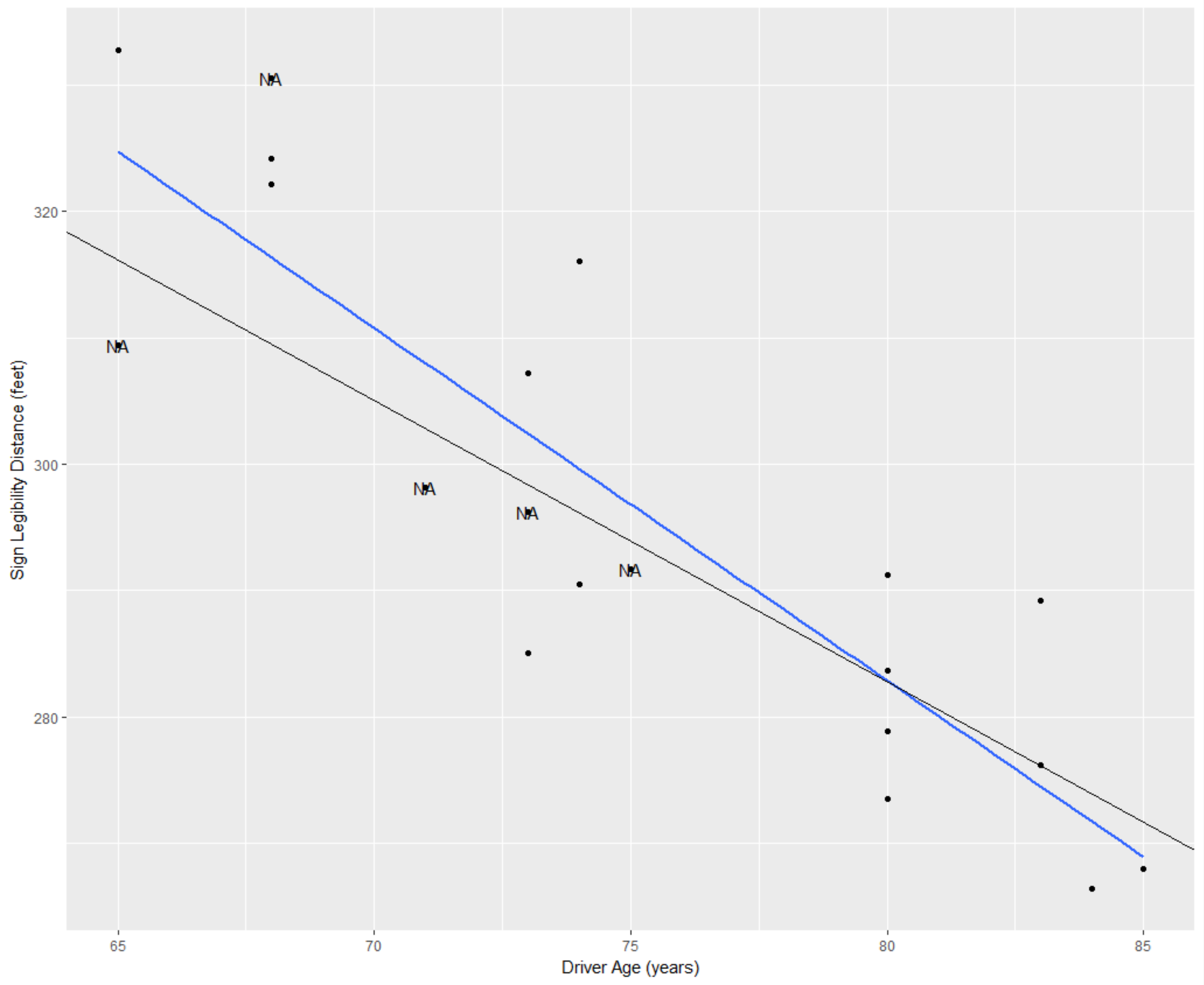
Mean imputation replaces every single missing value of a variable with the mean of the complete cases of the same variable

	age	distance
1	80	278.8375
2	83	289.1731
3	80	283.7063
4	83	276.2022
5	74	290.4947
6	68	324.1690
7	71	NA
8	75	NA
9	80	291.2071
10	85	267.9872
11	65	332.7962
12	68	NA
13	80	273.5567
14	65	NA
15	73	285.0229
16	74	316.0510
17	73	NA
18	73	307.2038
19	68	322.1212
20	84	266.3769

```
miss_meanimp<-miss
miss_meanimp[is.na(miss)]<-
mean(miss$distance,na.rm = TRUE)
```

	age	distance	r
1	80	278.8375	
2	83	289.1731	
3	80	283.7063	
4	83	276.2022	
5	74	290.4947	
6	68	324.1690	
7	71	293.6604	NA
8	75	293.6604	NA
9	80	291.2071	
10	85	267.9872	
11	65	332.7962	
12	68	293.6604	NA
13	80	273.5567	
14	65	293.6604	NA
15	73	285.0229	
16	74	316.0510	
17	73	293.6604	NA
18	73	307.2038	
19	68	322.1212	
20	84	266.3769	

```
> coef(lm(distance~age,data=miss_meanimp))
(Intercept)         age
   460.6903     -2.2241
```



MULTIPLE IMPUTATION

As the name suggests multiple imputation creates multiple imputed datasets based on different algorithms.

```
> library(mice)
> temp<-mice(data=miss, m=3, seed=12345)

 iter imp variable
  1   1  distance
  1   2  distance
  1   3  distance
  2   1  distance
  2   2  distance
  2   3  distance
  3   1  distance
  3   2  distance
  3   3  distance
  4   1  distance
  4   2  distance
  4   3  distance
  5   1  distance
  5   2  distance
  5   3  distance
```

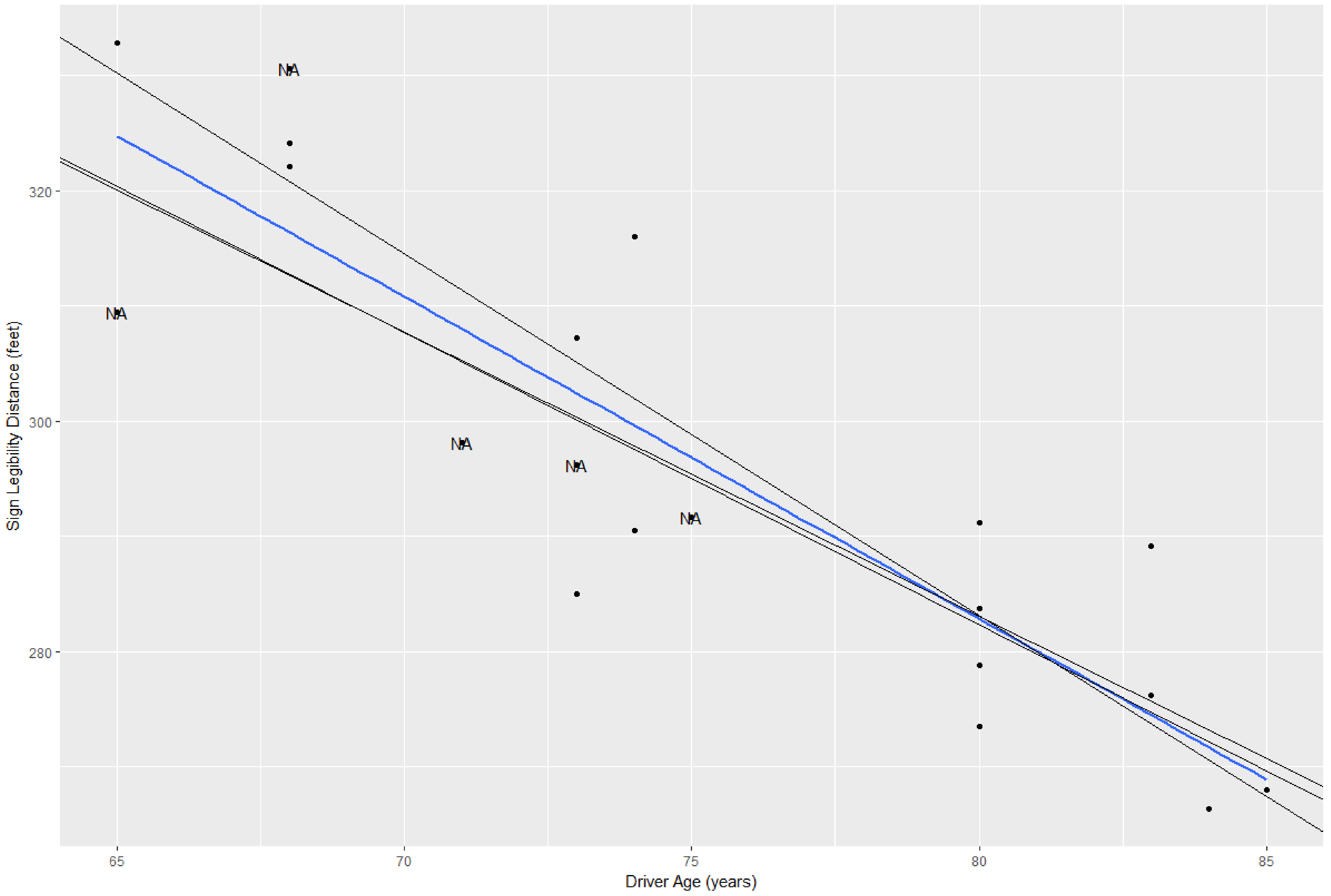
```
> temp$imp
$age
NULL

$distance
          1        2        3
7  290.4947 307.2038 316.0510
8  285.0229 285.0229 290.4947
12 285.0229 324.1690 322.1212
14 322.1212 332.7962 285.0229
17 307.2038 316.0510 285.0229

$r
NULL
```

```
> m1<-complete(temp,1)
> m2<-complete(temp,2)
> m3<-complete(temp,3)
```

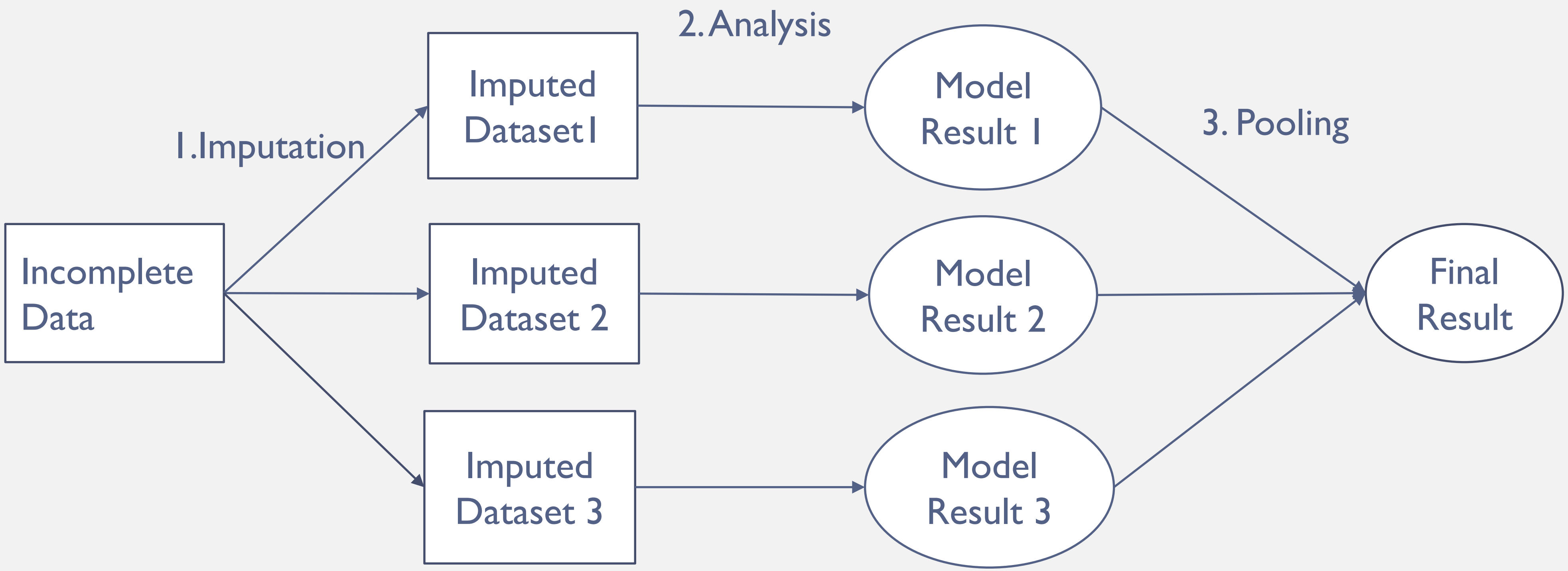
	age	distance	r
1	80	278.8375	
2	83	289.1731	
3	80	283.7063	
4	83	276.2022	
5	74	290.4947	
6	68	324.1690	
7	71	290.4947	NA
8	75	285.0229	NA
9	80	291.2071	
10	85	267.9872	
11	65	332.7962	
12	68	285.0229	NA
13	80	273.5567	
14	65	322.1212	NA
15	73	285.0229	
16	74	316.0510	
17	73	307.2038	NA
18	73	307.2038	
19	68	322.1212	
20	84	266.3769	

	age	distance	r
1	80	278.8375	
2	83	289.1731	
3	80	283.7063	
4	83	276.2022	
5	74	290.4947	
6	68	324.1690	
7	71	307.2038	NA
8	75	285.0229	NA
9	80	291.2071	
10	85	267.9872	
11	65	332.7962	
12	68	324.1690	NA
13	80	273.5567	
14	65	332.7962	NA
15	73	285.0229	
16	74	316.0510	
17	73	316.0510	NA
18	73	307.2038	
19	68	322.1212	
20	84	266.3769	

	age	distance	r
1	80	278.8375	
2	83	289.1731	
3	80	283.7063	
4	83	276.2022	
5	74	290.4947	
6	68	324.1690	
7	71	316.0510	NA
8	75	290.4947	NA
9	80	291.2071	
10	85	267.9872	
11	65	332.7962	
12	68	322.1212	NA
13	80	273.5567	
14	65	285.0229	NA
15	73	285.0229	
16	74	316.0510	
17	73	285.0229	NA
18	73	307.2038	
19	68	322.1212	
20	84	266.3769	



NA
NA
NA
NA
NA
320
300
280
65
70
75
80
85
Sign Legibility Distance (feet)
Driver Age (years)


Incomplete Data
1.Imputation
Imputed Dataset1
Imputed Dataset 2
Imputed Dataset 3
2. Analysis
Model Result 1
Model Result 2
Model Result 3
3. Pooling
Final Result

Pooling parameter estimates

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\widehat{Q_i}$$

Pooling standard errors

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i$$

$$B = \frac{1}{m}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$$

$$\sqrt{T} = \sqrt{\bar{U} + \left(1 + \frac{1}{m}\right)B}$$

Rubin (1987)

```
> mimodel<-with(temp,lm(distance~age))
> summary(pool(mimodel))
                   est          se         t        df
(Intercept) 499.568143 46.3729057 10.772845 3.516575
age          -2.708733  0.5938644 -4.561198 3.885096
                Pr(>|t|)        lo 95      hi 95    nmis
(Intercept) 0.0008122321  363.526358 635.60993    NA
age         0.0110559018  -4.376976  -1.04049     0
                  fmi      lambda
(Intercept) 0.6853859 0.5460706
age         0.6542114 0.5126425
```

ADDITIONAL MI PACKAGES

library(mice)

library (mi)

library(Amelia)

library(missForest)

library(Hmisc)

library(countimp)

MAXIMUM LIKELIHOOD

$$L = \prod_{i=1}^{m} f_i(y_{i1}, y_{i2}, \ldots., y_{ik}; Q) \prod_{m+1}^{n} f_i(y_{i3}, \ldots., y_{ik}; Q)$$

```
library(stats4)
mle()
library(lavaan)
sem(model,data,missing='fiml')
library(stats)
glm(model, family=poisson)
```

MI	ML
Better than traditional methods	Better than traditional methods
Can handle MCAR and MAR	Can handle MCAR and MAR
Multiple step needed to attain parameter estimates	Single step needed to attain parameter estimates
Not model specific	Model specific

TIPS

- Identifying the missing data mechanism can be hard. Talk to participants, other researchers to identify what causes missingness.
- Consider the percent of missingness and sample size.
- Consider the distribution of variables.
- Use a large number of imputations if using MI.
- Use both MI and ML if possible and see if you arrive at different conclusions.
- If possible simulate complete data that can mimic the scenario you are studying.

THANK YOU

bit.ly/MissingDataR

@MineDogucu
mdogucu
mdogucu@ncf.edu

REFERENCES

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198-1202.

Rubin, D.B B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.