Appendix 2 - R

Process data

Load data

In [1]:
```
# required to pre-load for piping
library('magrittr')

# set plot size
 options(plot.height=3, plot.width=3)

# load dodgers data file
er_data <- read.csv('ER_DATA.csv',sep='|')
```

Exploratory Data Analysis

In [2]:
```
# check data sample
head(er_data)
```

	AGE	SEX	RACE_ETHNICITY	PLAN_TYPE	STATE_CODE	PLAN_REGION	COMPLEXCARE_IND	MI
	<dbl>	<fct>	<fct>	<fct>	<fct>	<fct>	<int>	
1	38	F	White	MARKETPLACE	FL	SOUTHEAST	0	
2	81	M	White	MEDICAID	NY	NORTHEAST	1	
3	30	F	White	MARKETPLACE	TX	SOUTHWEST	0	
4	88	F	White	MEDICARE	TX	SOUTHWEST	0	
5	1	F	Hispanic	MEDICAID	NE	MIDDLESTATES	0	
6	59	M	Asian	MARKETPLACE	NY	NORTHEAST	0	

In [3]:
```
#View the structure of the data and the data types
str(er_data)
```

```
'data.frame':   69000 obs. of  46 variables:
 $ AGE                    : num  38 81 30 88 1 59 61 38 78 102 ...
 $ SEX                    : Factor w/ 2 levels "F","M": 1 2 1 1 1 2 2 2 2 1 ...
 $ RACE_ETHNICITY         : Factor w/ 7 levels "American Indian or A",..: 7 7 7 7 4 2 4
4 3 3 ...
 $ PLAN_TYPE              : Factor w/ 7 levels "BEHAVIORAL","COMMERCIAL",..: 5 6 5 7 6
5 6 6 7 6 ...
```

```
 $ STATE_CODE              : Factor w/ 37 levels "AL","AR","AZ",..: 7 27 34 34 22 27 34
34 8 4 ...
 $ PLAN_REGION             : Factor w/ 5 levels "MIDDLESTATES",..: 4 2 5 5 1 2 5 5 4 3
...
 $ COMPLEXCARE_IND         : int  0 1 0 0 0 0 1 0 0 1 ...
 $ MMP_DUAL_IND            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DUAL_PRODUCT_IND        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ LTC_IND                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MEDICAID_ELIGIBLE       : int  0 1 0 0 1 0 1 1 0 1 ...
 $ MEDICARE_ELIGIBLE       : int  0 0 0 1 0 0 0 0 2 0 ...
 $ BEHAVIORAL_ELIGIBLE     : int  0 0 0 0 0 0 1 1 0 0 ...
 $ COMMERCIAL_ELIGIBLE     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OTHER_ELIGIBLE          : int  1 0 1 0 0 1 0 0 0 0 ...
 $ RISK_TYPE_DESC          : Factor w/ 5 levels "","DUAL RISK",..: 1 1 1 1 1 1 1 1 1 3
...
 $ MEMBER_MONTHS_PRE       : num  12 12 1.94 4.93 12 ...
 $ ADD_STATE               : Factor w/ 73 levels "","12","13","15",..: 34 59 68 68 54 59
68 68 35 30 ...
 $ COUNTY_CLEAN            : Factor w/ 1221 levels "","  ","ABBEVILLE",..: 840 763 1190 1
973 750 354 648 1 642 ...
 $ REG_REGION_DESC         : Factor w/ 192 levels "** NO MATCH FOUND **",..: 5 2 22 2 12
4 2 52 91 2 2 ...
 $ RISK_SCORE              : num  2.744 3.18 0.46 0.58 0.249 ...
 $ PRIOR_TOTAL_COSTS_ANNUAL: num  2394 20920 417 521 1392 ...
 $ PRIOR_RX_COSTS_ANNUAL   : num  509.1 2091.12 416.59 433.81 3.63 ...
 $ ANNUAL_IP_COSTS         : num  0 13366 0 0 0 ...
 $ ANNUAL_ER_COSTS         : num  0 0 0 0 0 ...
 $ ANNUAL_OTHER_COSTS      : num  1885.4 5462.7 0 87.6 1388 ...
 $ FUTURE_RISK_INPATIENT   : num  1.267 5.192 0.617 2.255 0.7 ...
 $ BH_RISK_SCORE           : num  0.507 0.725 0.268 1.55 0.195 ...
 $ RX_RISK_SCORE           : num  1.518 5.802 0.636 3.03 0.424 ...
 $ ER_RISK_SCORE           : num  1.915 8.313 0.647 2.096 0.948 ...
 $ ORCA_SCORE              : int  93 2 1 17 0 83 32 4 87 1 ...
 $ ORCA_RISK_GROUP         : Factor w/ 4 levels "","HIGH","LOW",..: 4 3 3 3 3 4 3 3 4 3
...
 $ SUD_SEG_VALUE           : int  6 6 6 6 6 6 6 6 6 6 ...
 $ SUD_SEG_DEF             : Factor w/ 8 levels "","01: High Cost SUD Member - No Treatm
ent",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ ENG_SCORE               : num  40 78 56 90 38 77 83 26 66 96 ...
 $ POPHEALTHCAT_GROUPED    : Factor w/ 11 levels "01: Healthy",..: 6 5 1 3 1 4 9 1 9 1
...
 $ INTERVENABLE_IND        : int  0 1 0 0 0 0 0 0 0 0 ...
 $ SHORT_DESC              : Factor w/ 94 levels "","Acute and chronic renal failur
e",..: 64 36 68 56 64 36 77 23 63 23 ...
 $ SHORT_DESC_2            : Factor w/ 86 levels "Abdominal Infection/Pain",..: 23 44 35
59 23 44 64 22 45 22 ...
 $ RISK_CAT_RECODE         : Factor w/ 31 levels "","AIDS/HIV",..: 9 19 16 3 9 19 24 8 5
8 ...
 $ MEDICAID_CLAIMS         : int  0 1 0 0 1 0 1 0 0 1 ...
 $ MEDICARE_CLAIMS         : int  0 0 0 1 0 0 0 0 1 0 ...
 $ BEHAVIORAL_CLAIMS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ COMMERCIAL_CLAIMS       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OTHER_CLAIMS            : int  1 0 0 0 0 1 0 0 0 0 ...
 $ MORE_THAN_4_ER_VISITS   : int  0 0 0 0 0 0 0 0 0 0 ...
```

In [4]:

```
# Check data quality and basic statistics
psych::describe(er_data)
```

A psych: 46 x

	vars	n	mean	sd	median	trimmed
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

	vars	n	mean	sd	median	trimmed	
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
AGE	1	69000	3.202965e+01	2.226675e+01	29.0000	3.078391e+01	2
SEX*	2	69000	1.421623e+00	4.938224e-01	1.0000	1.402029e+00	
RACE_ETHNICITY*	3	69000	5.212435e+00	1.809818e+00	6.0000	5.327772e+00	
PLAN_TYPE*	4	69000	5.907913e+00	6.870157e-01	6.0000	5.969004e+00	
STATE_CODE*	5	69000	1.632291e+01	1.070050e+01	14.0000	1.567174e+01	1
PLAN_REGION*	6	69000	3.116449e+00	1.360601e+00	3.0000	3.145562e+00	
COMPLEXCARE_IND	7	69000	1.485217e-01	3.556190e-01	0.0000	6.065217e-02	
MMP_DUAL_IND	8	69000	1.802899e-02	1.330571e-01	0.0000	0.000000e+00	
DUAL_PRODUCT_IND	9	69000	2.978261e-02	1.699883e-01	0.0000	0.000000e+00	
LTC_IND	10	69000	1.128986e-02	1.056530e-01	0.0000	0.000000e+00	
MEDICAID_ELIGIBLE	11	69000	8.598116e-01	4.825234e-01	1.0000	8.802174e-01	
MEDICARE_ELIGIBLE	12	69000	9.028986e-02	2.895168e-01	0.0000	0.000000e+00	
BEHAVIORAL_ELIGIBLE	13	69000	2.700580e-01	4.524962e-01	0.0000	2.078442e-01	
COMMERCIAL_ELIGIBLE	14	69000	1.531884e-02	1.236418e-01	0.0000	0.000000e+00	
OTHER_ELIGIBLE	15	69000	1.138116e-01	3.223856e-01	0.0000	1.536232e-02	
RISK_TYPE_DESC*	16	69000	1.265072e+00	8.848860e-01	1.0000	1.000000e+00	
MEMBER_MONTHS_PRE	17	68998	9.855747e+00	3.584698e+00	12.0000	1.058881e+01	
ADD_STATE*	18	69000	4.341897e+01	1.565407e+01	40.0000	4.341953e+01	1
COUNTY_CLEAN*	19	69000	4.429531e+02	3.879715e+02	396.0000	4.138443e+02	58
REG_REGION_DESC*	20	69000	4.988143e+01	5.949340e+01	10.0000	4.119237e+01	1
RISK_SCORE	21	68935	3.293383e+00	7.210057e+00	1.0643	1.840817e+00	
PRIOR_TOTAL_COSTS_ANNUAL	22	68935	1.430178e+04	4.219008e+04	2511.1300	5.905007e+03	370
PRIOR_RX_COSTS_ANNUAL	23	68935	2.893580e+03	1.941766e+04	80.0000	4.996784e+02	11
ANNUAL_IP_COSTS	24	68935	3.286617e+03	1.975213e+04	0.0000	1.464175e+02	
ANNUAL_ER_COSTS	25	68935	8.081706e+02	2.291147e+03	121.9400	4.348911e+02	18
ANNUAL_OTHER_COSTS	26	68935	7.313408e+03	2.335150e+04	1309.2400	2.879420e+03	194
FUTURE_RISK_INPATIENT	27	68935	3.848280e+00	7.122357e+00	0.8823	1.850348e+00	
BH_RISK_SCORE	28	68935	5.133888e+00	1.093811e+01	0.7240	2.394595e+00	
RX_RISK_SCORE	29	68935	2.759930e+00	4.826812e+00	0.9749	1.740806e+00	
ER_RISK_SCORE	30	68935	5.087040e+00	7.156474e+00	1.5252	3.429224e+00	
ORCA_SCORE	31	65600	3.676759e+01	4.116319e+01	11.0000	3.355610e+01	1
ORCA_RISK_GROUP*	32	69000	2.876000e+00	6.863000e-01	3.0000	2.906594e+00	

	vars	n	mean	sd	median	trimmed	
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
SUD_SEG_VALUE	33	68935	5.449801e+00	1.321688e+00	6.0000	5.821393e+00	
SUD_SEG_DEF*	34	69000	6.543072e+00	1.175051e+00	7.0000	6.871123e+00	
ENG_SCORE	35	68935	5.111443e+01	3.022647e+01	51.0000	5.135368e+01	4
POPHEALTHCAT_GROUPED*	36	69000	3.672145e+00	2.472848e+00	3.0000	3.323841e+00	
INTERVENABLE_IND	37	69000	2.938986e-01	4.555493e-01	0.0000	2.423732e-01	
SHORT_DESC*	38	69000	5.014596e+01	2.633632e+01	60.0000	5.053147e+01	3
SHORT_DESC_2*	39	69000	4.187752e+01	2.439846e+01	35.0000	4.164670e+01	2
RISK_CAT_RECODE*	40	69000	1.481843e+01	9.223851e+00	14.0000	1.444676e+01	1
MEDICAID_CLAIMS	41	69000	6.745797e-01	4.685351e-01	1.0000	7.182246e-01	
MEDICARE_CLAIMS	42	69000	8.349275e-02	2.766276e-01	0.0000	0.000000e+00	
BEHAVIORAL_CLAIMS	43	69000	2.898551e-05	5.383780e-03	0.0000	0.000000e+00	
COMMERCIAL_CLAIMS	44	69000	1.223188e-02	1.099202e-01	0.0000	0.000000e+00	
OTHER_CLAIMS	45	69000	8.957971e-02	2.855808e-01	0.0000	0.000000e+00	
MORE_THAN_4_ER_VISITS	46	69000	4.637681e-01	4.986891e-01	0.0000	4.547101e-01	

Check categorical columns

In [5]:
```
# examine the data through descriptive statistics
Hmisc::describe(er_data)
```

```
er_data

 46  Variables      69000  Observations
--------------------------------------------------------------------------------
AGE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   69000        0      105        1    32.03    25.38        1        4
     .25      .50      .75      .90      .95
      14       29       49       63       71

lowest :   0   1   2   3   4, highest: 100 101 102 103 105
--------------------------------------------------------------------------------
SEX
       n  missing distinct
   69000        0        2

Value          F     M
Frequency  39908 29092
Proportion 0.578 0.422
--------------------------------------------------------------------------------
RACE_ETHNICITY
       n  missing distinct
   69000        0        7

lowest : American Indian or A Asian                Black or African Ame Hispanic
Native Hawaiian and
highest: Black or African Ame Hispanic             Native Hawaiian and  Unknown
```

```
White

American Indian or A (310, 0.004), Asian (2815, 0.041), Black or African Ame
(14842, 0.215), Hispanic (13209, 0.191), Native Hawaiian and (92, 0.001),
Unknown (8228, 0.119), White (29504, 0.428)
---------------------------------------------------------------------------
PLAN_TYPE
       n  missing distinct
   69000        0        7

lowest : BEHAVIORAL   COMMERCIAL   CORRECTIONAL DUALS        MARKETPLACE
highest: CORRECTIONAL DUALS        MARKETPLACE  MEDICAID     MEDICARE

Value        BEHAVIORAL   COMMERCIAL CORRECTIONAL        DUALS  MARKETPLACE
Frequency            10         1035          325           31         7210
Proportion        0.000        0.015        0.005        0.000        0.104

Value          MEDICAID     MEDICARE
Frequency         54306         6083
Proportion        0.787        0.088
---------------------------------------------------------------------------
STATE_CODE
       n  missing distinct
   69000        0       37

lowest : AL AR AZ CA CT, highest: TN TX VT WA WI
---------------------------------------------------------------------------
PLAN_REGION
       n  missing distinct
   69000        0        5

lowest : MIDDLESTATES NORTHEAST    PACIFIC      SOUTHEAST    SOUTHWEST
highest: MIDDLESTATES NORTHEAST    PACIFIC      SOUTHEAST    SOUTHWEST

Value      MIDDLESTATES    NORTHEAST      PACIFIC    SOUTHEAST    SOUTHWEST
Frequency         13600         8992        12706        23177        10525
Proportion        0.197        0.130        0.184        0.336        0.153
---------------------------------------------------------------------------
COMPLEXCARE_IND
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.379    10248   0.1485   0.2529

---------------------------------------------------------------------------
MMP_DUAL_IND
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.053     1244  0.01803  0.03541

---------------------------------------------------------------------------
DUAL_PRODUCT_IND
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.087     2055  0.02978  0.05779

---------------------------------------------------------------------------
LTC_IND
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.033      779  0.01129  0.02233

---------------------------------------------------------------------------
MEDICAID_ELIGIBLE
       n  missing distinct     Info     Mean      Gmd
   69000        0        5    0.572   0.8598   0.4201

lowest : 0 1 2 3 5, highest: 0 1 2 3 5

Value          0     1     2     3     5
```

```
Frequency  13512 51681  3777    29     1
Proportion 0.196 0.749 0.055 0.000 0.000
--------------------------------------------------------------------------------
MEDICARE_ELIGIBLE
       n  missing distinct     Info     Mean      Gmd
   69000        0        3    0.244  0.09029   0.1646

Value          0     1     2
Frequency  62828  6114    58
Proportion 0.911 0.089 0.001
--------------------------------------------------------------------------------
BEHAVIORAL_ELIGIBLE
       n  missing distinct     Info     Mean      Gmd
   69000        0        4    0.587   0.2701   0.3983

Value          0     1     2     3
Frequency  50627 18114   257     2
Proportion 0.734 0.263 0.004 0.000
--------------------------------------------------------------------------------
COMMERCIAL_ELIGIBLE
       n  missing distinct     Info     Mean      Gmd
   69000        0        3    0.045  0.01532  0.03017

Value          0     1     2
Frequency  67950  1043     7
Proportion 0.985 0.015 0.000
--------------------------------------------------------------------------------
OTHER_ELIGIBLE
       n  missing distinct     Info     Mean      Gmd
   69000        0        4    0.299   0.1138   0.2024

Value          0     1     2     3
Frequency  61252  7644   103     1
Proportion 0.888 0.111 0.001 0.000
--------------------------------------------------------------------------------
RISK_TYPE_DESC
       n  missing distinct
   69000        0        5

lowest :                 DUAL RISK       FEE FOR SERVICE FULL RISK       SHARED RISK
highest:                 DUAL RISK       FEE FOR SERVICE FULL RISK       SHARED RISK

Value                           DUAL RISK FEE FOR SERVICE       FULL RISK
Frequency            62109            1696            1681             824
Proportion           0.900           0.025           0.024           0.012

Value          SHARED RISK
Frequency             2690
Proportion           0.039
--------------------------------------------------------------------------------
MEMBER_MONTHS_PRE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68998        2      362    0.707    9.856    3.278    1.941    2.928
     .25      .50      .75      .90      .95
   7.895   12.000   12.000   12.000   12.000

lowest :  0.0658  0.0987  0.1316  0.1645  0.1974
highest: 11.8750 11.9079 11.9408 11.9737 12.0000
--------------------------------------------------------------------------------
ADD_STATE
       n  missing distinct
   69000        0       73

lowest :    12 13 15 17, highest: UT VA VT WA WI
--------------------------------------------------------------------------------
```

```
COUNTY_CLEAN
       n  missing distinct
   69000        0     1221

lowest :                     ABBEVILLE ACADIA    ACCOMACK
highest: YOUNG     YUBA      YUMA      ZAPATA    ZAVALA
------------------------------------------------------------------------------
REG_REGION_DESC
       n  missing distinct
   69000        0      192

lowest : ** NO MATCH FOUND **       ** NOT PROVIDED   **       AD
All AZ Regions             All FL Regions
highest: WEST CENTRAL INDIANA       West CFC - Cinci Child     West CFC - Non Cinci Chil
d WESTERN                   Wisconsin
------------------------------------------------------------------------------
RISK_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    35194        1    3.293    4.595   0.1000   0.1358
     .25      .50      .75      .90      .95
  0.3543   1.0643   3.4656   7.6434  11.9494

lowest :   0.1000   0.1016   0.1074   0.1155   0.1241
highest:  99.1961  99.6001 100.3859 101.8011 105.0436
------------------------------------------------------------------------------
PRIOR_TOTAL_COSTS_ANNUAL
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    58730    0.999    14302    22748      0.0      0.0
     .25      .50      .75      .90      .95
   387.7   2511.1  10858.3  35598.6  65995.7

lowest :       0.00       0.15       0.20       0.24       0.27
highest: 1272214.85 1664410.65 1724977.95 2105454.98 2116494.01
------------------------------------------------------------------------------
PRIOR_RX_COSTS_ANNUAL
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    37449    0.981     2894     5265      0.0      0.0
     .25      .50      .75      .90      .95
     0.0     80.0    763.1   5007.3  11636.1

lowest :       0.00       0.01       0.02       0.03       0.04
highest:  834213.58  913792.63 1203595.10 2076825.97 2103466.22
------------------------------------------------------------------------------
ANNUAL_IP_COSTS
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65     9017    0.382     3287     6243        0        0
     .25      .50      .75      .90      .95
       0        0        0     4695    15488

lowest :       0.00       4.82      13.73      34.24      50.22
highest:  728725.33  804726.76  844176.59 1335664.48 1452105.77
------------------------------------------------------------------------------
ANNUAL_ER_COSTS
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    32405    0.896    808.2     1235      0.0      0.0
     .25      .50      .75      .90      .95
     0.0    121.9    989.0   2199.9   3349.0

lowest :      0.00      0.24      0.30      1.24      1.27
highest:  75902.95  81748.40  97316.47 117117.70 312528.31

Value           0   5000  10000  15000  20000  25000  30000  35000  40000
Frequency   63306   4929    491    100     55     14     10      9      8
Proportion  0.918  0.072  0.007  0.001  0.001  0.000  0.000  0.000  0.000
```

```
Value       45000  55000  60000  70000  75000  80000  95000 115000 315000
Frequency       3      1      1      3      1      1      1      1      1
Proportion  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000

For the frequency table, variable is rounded to the nearest 5000
--------------------------------------------------------------------------------
ANNUAL_OTHER_COSTS
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    54845    0.996     7313    11973      0.0      0.0
     .25      .50      .75      .90      .95
   206.8   1309.2   5421.9  16276.1  32059.1

lowest : -418331.1 -337918.3 -187334.9 -159611.5 -149616.9
highest:  594933.1  636623.9  660672.7  994105.1 1239582.3
--------------------------------------------------------------------------------
FUTURE_RISK_INPATIENT
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    24523        1    3.848     5.31   0.5879   0.6075
     .25      .50      .75      .90      .95
  0.6386   0.8823   2.5404  11.9109  22.3316

lowest :  0.5858  0.5879  0.5910  0.5932  0.5939
highest: 36.6167 36.6188 36.6259 36.6287 36.6300
--------------------------------------------------------------------------------
BH_RISK_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    16146    0.999    5.134    7.887    0.131    0.195
     .25      .50      .75      .90      .95
   0.270    0.724    4.239   15.304   26.658

lowest :   0.100   0.101   0.102   0.103   0.104
highest: 130.419 130.538 133.453 135.587 139.763
--------------------------------------------------------------------------------
RX_RISK_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    28740        1     2.76    3.555   0.2046   0.2852
     .25      .50      .75      .90      .95
  0.4717   0.9749   3.1694   6.9881  10.1655

lowest :  0.1347  0.1742  0.1794  0.1816  0.1846
highest: 63.8118 65.6683 66.6049 71.9618 74.5865
--------------------------------------------------------------------------------
ER_RISK_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65    28667        1    5.087    6.586   0.3620   0.4464
     .25      .50      .75      .90      .95
  0.6667   1.5252   6.2923  16.6256  22.4544

lowest :  0.2896  0.3320  0.3374  0.3469  0.3473
highest: 33.5977 33.6080 33.6198 33.6532 33.7343
--------------------------------------------------------------------------------
ORCA_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   65600     3400       99    0.943    36.77    43.96        0        0
     .25      .50      .75      .90      .95
       0       11       85       98       99

lowest :   0   1   2   3   4, highest:  96  97  98  99 100
--------------------------------------------------------------------------------
ORCA_RISK_GROUP
       n  missing distinct
   69000        0        4

Value              HIGH    LOW MEDIUM
Frequency     3400  10858  45640   9102
```

```
Proportion  0.049  0.157  0.661  0.132
--------------------------------------------------------------------------------
SUD_SEG_VALUE
       n  missing distinct     Info     Mean      Gmd
   68935       65        7    0.483     5.45   0.9522

lowest : 0 1 2 3 4, highest: 2 3 4 5 6

Value          0     1     2     3     4     5     6
Frequency    970  1284  1921  4232    67  5174 55287
Proportion 0.014 0.019 0.028 0.061 0.001 0.075 0.802
--------------------------------------------------------------------------------
SUD_SEG_DEF
       n  missing distinct
   69000        0        8

lowest :                                            01: High Cost SUD Member - No Treatme
nt   02: High Cost SUD Member - Some Treatment 03: Not High Cost SUD Member
04: Harmful Use Rx Only (No SUD Dx)
highest: 03: Not High Cost SUD Member              04: Harmful Use Rx Only (No SUD Dx)
05: Nicotine Only (Not in segments 1-4)   06: No known SUD behavior                 New
--------------------------------------------------------------------------------
ENG_SCORE
       n  missing distinct     Info     Mean      Gmd      .05      .10
   68935       65      101        1    51.11    34.88        4        9
     .25      .50      .75      .90      .95
      24       51       79       92       96

lowest :   0   1   2   3   4, highest:  96  97  98  99 100
--------------------------------------------------------------------------------
POPHEALTHCAT_GROUPED
       n  missing distinct
   69000        0       11

lowest : 01: Healthy                          02: Acute Episodic                  03: Chr
onic Stable PH/BH            04: Health Coaching                 05: Chronic Interventio
nal PH/BH CM
highest: 07: Catastrophic Conditions         08: Dementia and Custodial Care     09: LTS
S                            10: End of Life Care                99: Unclassified
--------------------------------------------------------------------------------
INTERVENABLE_IND
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.623    20279   0.2939   0.4151

--------------------------------------------------------------------------------
SHORT_DESC
       n  missing distinct
   69000        0       94

lowest :                                              Acute and chronic renal failure
Acute bronchitis                                Agents used to treat enzyme deficiency sta
tes AIDS/HIV
highest: Septicemia                                   Sickle-cell anemia
Substance Abuse                                 Ulcers, gastritis/duodenitis
Valvular disorders
--------------------------------------------------------------------------------
SHORT_DESC_2
       n  missing distinct
   69000        0       86

lowest : Abdominal Infection/Pain        Acute And Chronic Renal Failure Acute Bronchiti
s               Adhd/Idd/Autism                 Aids/Hiv
highest: Substance Abuse                 Ulcers, Gastritis/Duodenitis    Upper Gi Inflam
mation/Infection Urology                        Valvular Disorders
--------------------------------------------------------------------------------
```

```
RISK_CAT_RECODE
       n  missing distinct
   69000        0       31

lowest :                              AIDS/HIV                     Behavioral Health (In
c. SUD) Cancer                       Cardiology
highest: Orthopedics                  Other Medical                Pulmonology
Rheumatology                Urology
--------------------------------------------------------------------------------
MEDICAID_CLAIMS
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.659    46546   0.6746   0.4391

--------------------------------------------------------------------------------
MEDICARE_CLAIMS
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2     0.23     5761  0.08349    0.153

--------------------------------------------------------------------------------
BEHAVIORAL_CLAIMS
        n   missing  distinct      Info       Sum      Mean       Gmd
    69000         0         2         0         2 2.899e-05 5.797e-05

--------------------------------------------------------------------------------
COMMERCIAL_CLAIMS
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.036      844  0.01223  0.02416

--------------------------------------------------------------------------------
OTHER_CLAIMS
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.245     6181  0.08958   0.1631

--------------------------------------------------------------------------------
MORE_THAN_4_ER_VISITS
       n  missing distinct     Info      Sum     Mean      Gmd
   69000        0        2    0.746    32000   0.4638   0.4974

--------------------------------------------------------------------------------
```

Therefore, the observations are:

- The outcome variable is fairly balanced, as such there will not be a need to oversample or downsample the data
- The data is fairly clean with minimimal missing records.
- The majority of patients are female at fifthy eight percent compared to male are forthy two percent
- The data is made up of sixty nine thousand records and forthy six variables

Box plot

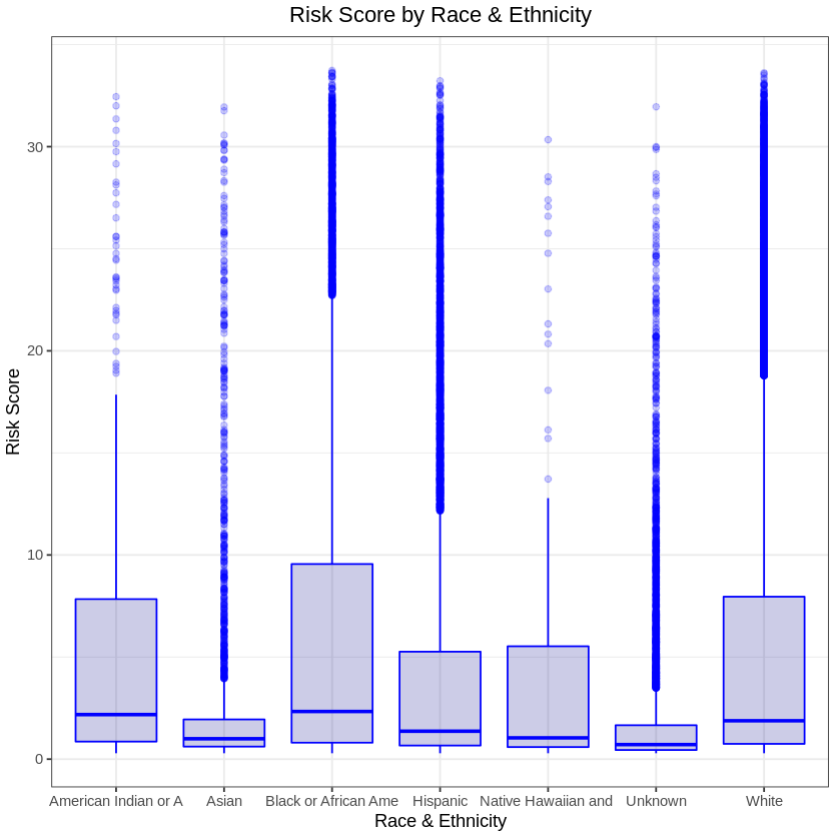
ER Risk Score by Race & Ethnicity

In [6]:
```
# set plot size
 options(plot.height=3, plot.width=3)
```

In [7]:

```
ggplot2::ggplot(er_data, ggplot2::aes(x=RACE_ETHNICITY, y=ER_RISK_SCORE)) +
    ggplot2::geom_boxplot(color="blue", fill="darkblue", alpha=0.2) +
    ggplot2::labs(title='Risk Score by Race & Ethnicity',
                  x='Race & Ethnicity',
                  y='Risk Score')+
    ggplot2::theme_bw() +
    ggplot2::theme(plot.title = ggplot2::element_text(hjust = 0.5))
```

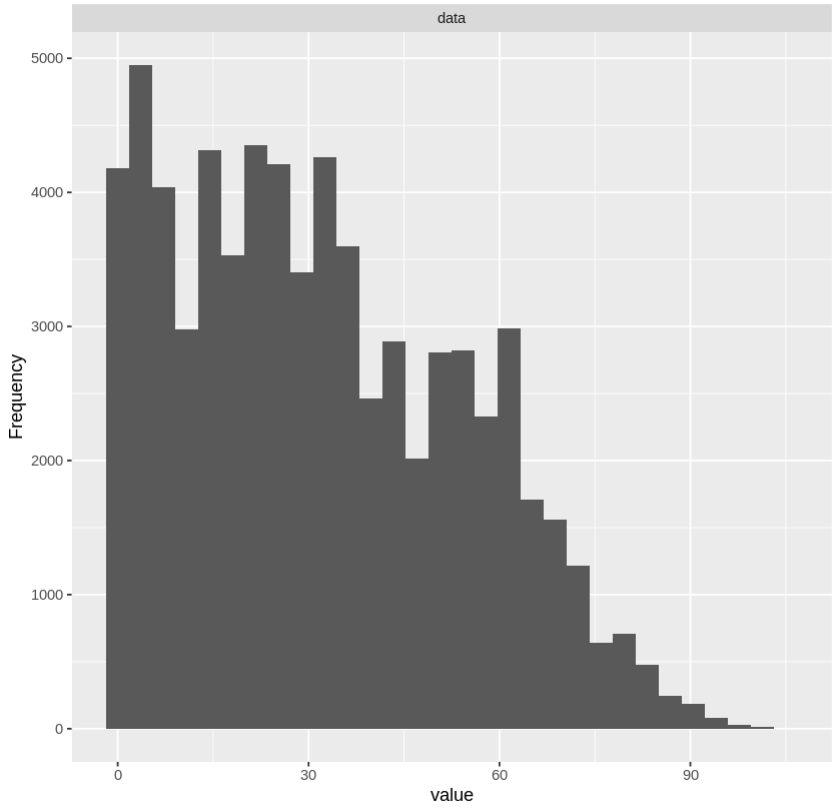
```
Warning message:
“Removed 65 rows containing non-finite values (stat_boxplot).”
```



Based on the plot, african american patients seem to have higher risk scores compared to other race & ethnicity

In [8]:

```
# Histogram of the AGE variable
DataExplorer::plot_histogram(er_data$AGE)
```

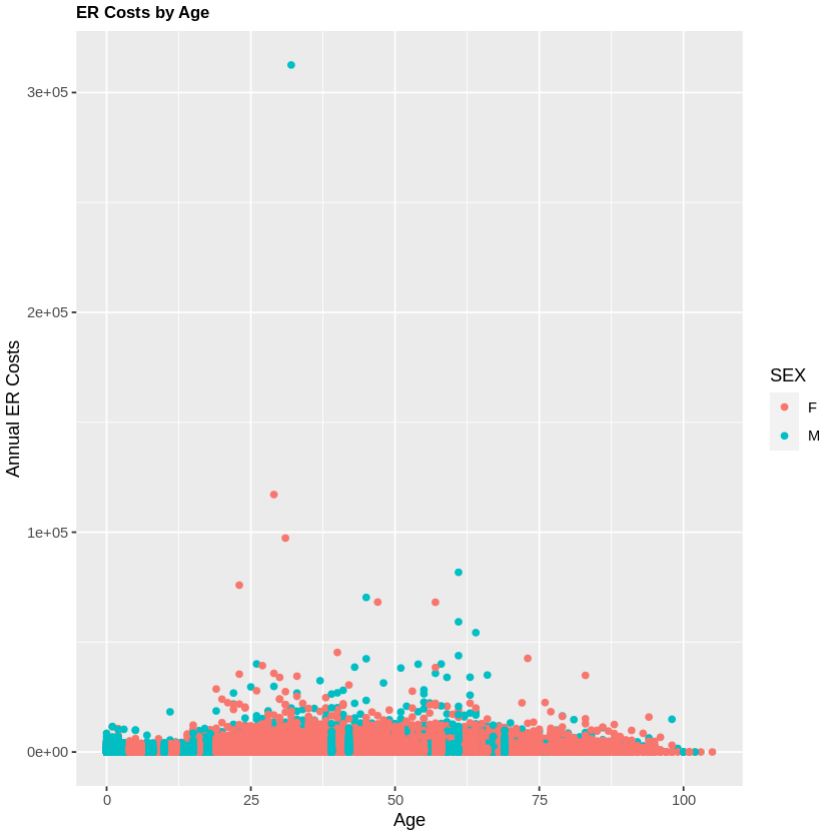


The data is highly skewed and patients are relatively younger with the majority of the population being 80 years old or younger as you would expect.

In [9]:

```
#Strip Plot of Attendance by opponent or visiting team
ggplot2::ggplot(er_data, ggplot2::aes(x=AGE, y=ANNUAL_ER_COSTS, color = SEX)) +
        ggplot2::geom_point() +
        ggplot2::labs(title = 'ER Costs by Age',
                      x='Age',
                      y='Annual ER Costs') +
        ggplot2::theme(plot.title = ggplot2::element_text(lineheight=3, face="bold", co
```

```
Warning message:
“Removed 65 rows containing missing values (geom_point).”
```



This plot further confirms that the data is fairly contained with relatively very few outliers.