ANA 610 Homework #4

Fortune Corp, a maker of specialized laboratory equipment for the pharmaceutical industry, began business in June 1980. Priding itself on employee job satisfaction, the company is seeking to understand why employees voluntarily leave the company.

Over the last 3 years, at the request of the SVP of Human Resources, the HR department has been conducting an employee survey. The SVP wants enough data collected so that a predictive model of employee voluntary attrition can be built and tested. The objective is to use such a model to find current employees who might be thinking of leaving, so proactive steps can be taken to retain them.

The SVP now thinks there has been enough data collected. So, she has requested that you, the lead data science team, take over and work your magic!

The qualifications for the target sample are having taken the survey. This sample is broken into two segments:

1. Employees who voluntarily attritioned (left the company)
2. Employees who are still with the company

Assume the analysis is taking place June 1, 2018.

The following 5 data files have been created for your use by the IT department:

- (csv) Credit Bureau file: fortune_credit.csv
    - FICO score (SVP thinks this might be predictive)
- (SAS) Accounting file: fortune_acct
    - Payroll data
- (SAS) Attrition file: fortune_attrition
    - Employees who have left the company over the 2015-2017 period
- (SAS) HR file fortune_hr
    - Background employee data
- (SAS) Survey file fortune_survey
    - Data collected from the employee survey

Unfortunately, there is no data-dictionary for these files. But most fields should be self-explanatory.

The data files are available online at SAS Studio.

Task #1 (100 pts): Generate a data audit report (using the audit report template) to be shared with both the HR and IT department; include a check of the available modeling sample size. Assemble all 5 data files into a single, modeling dataset.

Task #2 (20 pts):

1. Deduplicate your modeling dataset:
    a. Show your SAS code

```
/* TASK 2, QUESTION 1 */
/* Check for duplicates */
*====> raw nobs;
proc sql; select count(*) into : nobs from ana610.fortune_master; quit;
*====> deduped;
proc sort data=ana610.fortune_master out=ana610.fortune_master_clean nodupkey; by employee_no; run;
proc sql; select count(*) into : nobs from ana610.fortune_master_clean; quit;
/* 25 duplicate employee_no identified */
```

b. Show the before and after observation (row) count

Before:

4892

After (25 duplicate observations identified):

4867

2. **Using your deduplicated dataset**, create two variables, one for AGE, employee age (in years), and one for TENURE, the length of time the employee has been with the company (in years). Assume on average each year has 365.25 days. Create AGE and TENURE for **all** employees in the dataset. **HINT**: how should TENURE be defined for those who left the company? Show your SAS code.

```
/* TASK 2, QUESTION 2 */
data work.fortune_master_clean; set ana610.fortune_master_clean;
    /* Age variable */
    age_day = mdy(6,1,2018) - birth_dt;
    age = round(age_day/365.25);
    /* Tenure variable */
    if missing(depart_dt) then tenure_day = mdy(6,1,2018)- hire_dt;
    else tenure_day = depart_dt - hire_dt;
    tenure = round(tenure_day/365.25);
run;
```

3. Using PROC UNIVARIATE, check AGE and TENURE for integrity issues. Specifically check for (a) missing values; (b) extreme values; and (c) extreme distribution. Discuss your findings. Show the relevant PROC UNIVARIATE charts and tables.

a) Missing values:

Age is missing 270 values. This is because the Age variable was created by using "birth_dt" in Task 2, Question 2. Since "birth_dt" already had 270 missing values, Age will also have 270 missing values.	**Tenure has 0 missing values. This is because I created the Tenure variable using the "depart_dt" and "hire_dt" variables in Task 2, Question 2. Tenure, for an active employee, is defined as the period of time between our analysis date (6/1/2018) and their hire date. Tenure for an inactive employee is defined as the period of time between their depart date and their hire date. Since all employees are**

(table below)	covered, Tenure should not have any missing values.

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	270	5.55	100.00

Analysis Variable : tenure					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	0	8.3688104	7.0000000	42.0000000

b) Extreme values:

Age:
Using a Top/Bottom approach, the 1-99% of the data falls between 21-60. PROC UNIVARIATE identifies 19 as a low extreme and 61-62 high extreme. Comparing these values to the range of this dataset (19-62), I do not think there are extreme values for the Age variable. This is because the distribution of values are representative of the data.

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
19	4601	61	4491
19	4404	61	4726
19	3806	61	4819
19	3201	62	189
19	2319	62	2739
19	2123	62	2980
19	1806	62	3579
19	1315	62	3843
19	1095	62	4080
19	850	62	4750

Quantiles (Definition 5)	
Level	Quantile
100% Max	62
99%	60
95%	56
90%	52
75% Q3	44
50% Median	37
25% Q1	32
10%	28
5%	25
1%	21
0% Min	19

Tenure
Using a Top/Bottom approach, the 1-99% of the data falls between 1-31. PROC UNIVARIATE identifies 0 as a low extreme and 38-42 high extreme. The company started in June 1, 1980, so the highest amount of Tenure an employee could have is 38 years (analysis date June 1, 2018). Since the tenure variable was created using hire_dt, any employee data that was hired before June 1, 1980 should be considered outliers (Count: 6 obs).

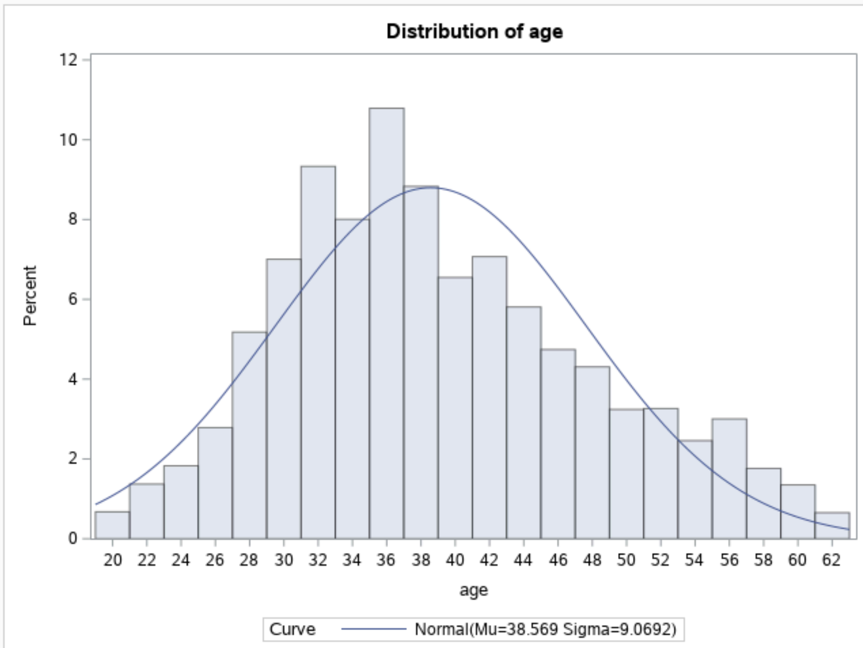
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	4848	38	351
0	4644	38	924
0	4429	38	2322
0	4212	38	2725
0	4073	38	2915
0	3356	38	4127
0	3054	39	1149
0	2944	39	2933
0	2653	40	706
0	2123	42	992

Quantiles (Definition 5)	
Level	Quantile
100% Max	42
99%	31
95%	22
90%	16
75% Q3	11
50% Median	7
25% Q1	4
10%	3
5%	2
1%	1
0% Min	0

c) Extreme distribution:
Age: The distribution for Age appears to be normal with a low, positive skewness value (0.421). Since this skewness value is lower than 5, this distribution is not considered extreme.

The UNIVARIATE Procedure
Variable: age

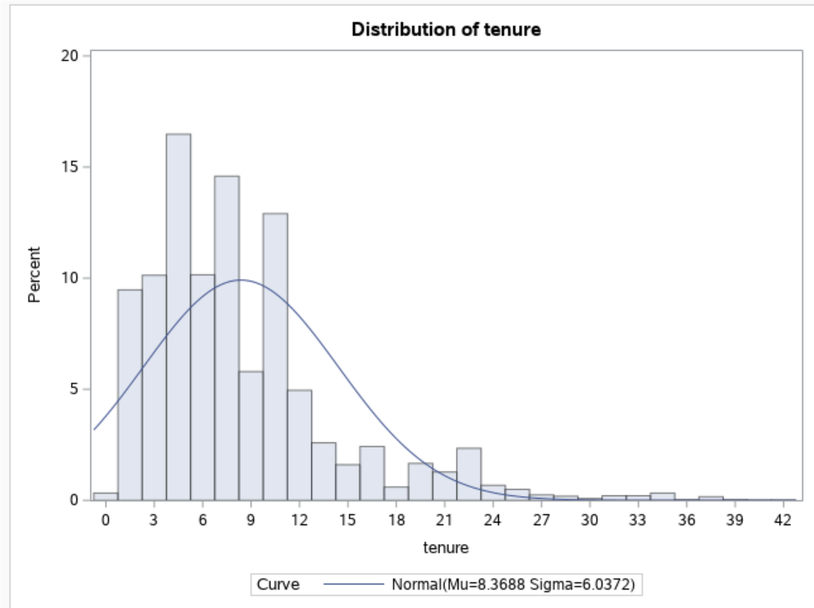
Moments			
N	4597	Sum Weights	4597
Mean	38.5688492	Sum Observations	177301
Std Deviation	9.06923206	Variance	82.2509702
Skewness	0.42097246	Kurtosis	-0.3956512
Uncorrected SS	7216321	Corrected SS	378025.459
Coeff Variation	23.5143963	Std Error Mean	0.13376216



Tenure: The distribution for Tenure appears to be skewed right with a skewness value of (1.682). Since this skewness value is lower than 5, this distribution is not considered extreme.

The UNIVARIATE Procedure
Variable: tenure

Moments			
N	4867	Sum Weights	4867
Mean	8.36881036	Sum Observations	40731
Std Deviation	6.03716714	Variance	36.4473871
Skewness	1.68192383	Kurtosis	3.69610767
Uncorrected SS	518223	Corrected SS	177352.985
Coeff Variation	72.1388929	Std Error Mean	0.08653714



4. **Using your deduplicated dataset**, create a target variable, ATT_Q, which takes on a value of 1 if an employee took the survey and voluntarily attritioned; or 0 if the employee took the survey and did not attrition. Show the relevant SAS output from PROC FREQ which shows how many employees fall in each segment.

att_q	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1233	83.88	1233	83.88
1	237	16.12	1470	100.00
Frequency Missing = 3397				

Task #3 (80 pts):

Note: Use the deduplicated modeling dataset you created in Task #2.

Using PROC HPBIN and target-based enumeration,

1. Fill in the following table template to analyze how many bins are appropriate for AGE and TENURE. Start with 10 bins. Then, using a threshold of 20 for <u>each segment</u> of the target variable, determine how many bins should be created (HINT: 10 is too many). If there are missing values, ignore the bin that captures these in your final bin count.
   For each final bin count, I chose the bin size that yielded the most of number of bins that passed the threshold and also the best Information Value (IV). Ideally, I want my IV value to be between 0.3-0.5 to indicate a strong predictor power. I chose the bolded items in the table below for my final bin count.

   < 0.1 = Weak
   0.1 - 0.3 = Medium
   0.3 - 0.5 = Strong

0.5+ = Suspicious

For the Tenure variable, using the Equal Width binning type, Bins 1-10 were tested to see if all bins could meet the 20 segment threshold. However, after testing Bins 1-10, no bin size was able to have all the bins meet the 20 segment threshold. A bin size of 6 yielded the most number of bins that passed the threshold and also had the best IV value.

Binning Type		Number of Bins	Number of Bins which Pass Threshold	Variable-level Information Value (IV)
Equal Height	Starting bin count	Age: 10 Tenure: 10	Age: 4 Tenure: 3	Age: 0.283 Tenure: 0.946
	Final bin count	Age: 6 **Age: 5** Age: 4 Tenure: 3 **Tenure: 2**	Age: 6 **Age: 5** Age: 4 Tenure: 3 **Tenure: 2**	Age: 0.262 **Age: 0.265** Age: 0.250 Tenure: 0.567 **Tenure: 0.260**
Equal Width	Starting bin count	Age: 10 Tenure: 10	Age: 5 Tenure: 3	Age: 0.309 Tenure: 0.591
	Final bin count	**Age: 3** Age: 2 Tenure: 9 **Tenure: 6** Tenure: 4	**Age: 3** Age: 2 Tenure: 3 **Tenure: 2** Tenure: 1	**Age: 0.217** Age: 0.105 Tenure: 0.581 **Tenure: 0.388** Tenure: 0.340

2. For each variable AGE and Tenure, support your final bin count by showing the relevant PROC HPBIN output which displays the counts by bin/segment.
Equal Height / Age / Final Bin Count: 5

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	BinsMeta
Number of Bins Specified	See BinsMeta
Number of Variables	1

Number of Observations Read	4867
Number of Observations Used	1395

Weight of Evidence

Variable	Binned Variable	Range	Non-event Count	Non-event Rate	Event Count	Event Rate	Weight of Evidence	Information Value
age	BIN_age		61	0.81333333	14	0.18666667	-0.1773288	0.00170216
		age < 31.0013	240	0.72948328	89	0.27051672	-0.6571428	0.11886414
		31.0013 <= age < 35.0003	195	0.79918033	49	0.20081967	-0.2679661	0.01302321
		35.0003 <= age < 40.0012	274	0.89836066	31	0.10163934	0.52999554	0.04845248
		40.0012 <= age < 47.0016	246	0.91111111	24	0.08888889	0.67813234	0.06662485
		47.0016 <= age	217	0.87854251	30	0.12145749	0.32955461	0.01628370

Variable Information Value	
Variable	Information Value
age	0.26495053

Equal Height / Tenure / Final Bin Count: 2

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	BinsMeta
Number of Bins Specified	See BinsMeta
Number of Variables	1

Number of Observations Read	4867
Number of Observations Used	1470

Weight of Evidence

Variable	Binned Variable	Range	Non-event Count	Non-event Rate	Event Count	Event Rate	Weight of Evidence	Information Value
tenure	BIN_tenure		0	0	0	0	0	0
		tenure < 7.0014	651	0.78151261	182	0.21848739	-0.3746424	0.08989618
		7.0014 <= tenure	582	0.91365777	55	0.08634223	0.70999190	0.17036394

Variable Information Value	
Variable	Information Value
tenure	0.26026012

Equal Width / Age / Final Bin Count: 3

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	BinsMeta
Number of Bins Specified	See BinsMeta
Number of Variables	1

Number of Observations Read	4867
Number of Observations Used	1395

Weight of Evidence

Variable	Binned Variable	Range	Non-event Count	Non-event Rate	Event Count	Event Rate	Weight of Evidence	Information Value
age	BIN_age		61	0.81333333	14	0.18666667	-0.1773288	0.00170216
		age < 33.333333333	339	0.74505495	116	0.25494505	-0.5767354	0.12371685
		33.333333333 <= age < 47.666666667	616	0.88888889	77	0.11111111	0.43029618	0.07517273
		47.666666667 <= age	217	0.87854251	30	0.12145749	0.32955461	0.01628370

Variable Information Value	
Variable	Information Value
age	0.21687544

Equal Width / Tenure / Final Bin Count: 6

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	BinsMeta
Number of Bins Specified	See BinsMeta
Number of Variables	1

Number of Observations Read	4867
Number of Observations Used	1470

Weight of Evidence								
Variable	Binned Variable	Range	Non-event Count	Non-event Rate	Event Count	Event Rate	Weight of Evidence	Information Value
tenure	BIN_tenure		0	0	0	0	0	0
		tenure < 7	526	0.75466284	171	0.24533716	-0.5255077	0.15498126
		7 <= tenure < 14	520	0.91228070	50	0.08771930	0.69266044	0.14598868
		14 <= tenure < 21	101	0.92660550	8	0.07339450	0.88653361	0.04269436
		21 <= tenure < 28	67	0.94366197	4	0.05633803	1.16925290	0.04380182
		28 <= tenure < 35	13	0.81250000	3	0.18750000	-0.1828083	0.00038661
		35 <= tenure	6	0.85714286	1	0.14285714	0.14261411	0.00009224

Variable Information Value	
Variable	Information Value
tenure	0.38794496

3. For each variable AGE and TENURE, explain, if you had to pick one type of binning, which binning type (equal height or equal width) should be selected for your predictive model of attrition. Focus on the variable-level IV to support your selection.
I would choose equal height as my binning type for both variables, Age and Tenure. In the final bin count, the variable-level IV for Age and Tenure is 0.265 and 0.260, respectively. Both of these IV values are considered a "Medium" predictive power. Equal height's "Medium" predictive powers were also on the higher side of the "Medium" tier (closer to 0.3, which is the beginning of the "Strong" predictive tier).

< 0.1 = Weak
0.1 - 0.3 = Medium
0.3 - 0.5 = Strong
0.5+ = Suspicious

Equal width, in comparison, has a similar predictive power for Age and Tenure, and it produced a variable-level IV for Age and Tenure is 0.217 and 0.388, respectively. The Age IV value would be a considered a "Medium" predictive power and the "Tenure" predictive power would be considered a "Strong" predictive power. However, in the final bin count for Tenure, all the bins could not cohesively meet the 20 segment threshold.

4. For each variable AGE and TENURE, create dummy variables for each bin for your selected binning type. The dummy variable names should show the relevant bin ranges. Using PROC MEANS and PROC TRANSPOSE, produce a “tall and skinny” output table for each variable which shows the bin dummy variables as rows with the sum for each bin by target variable segment as columns (so, N rows by 3 columns). Check that for each bin/segment the threshold is met and that the sum for each bin matches the output from your final PROC HPBIN run.

Dummy variables for Age and Tenure was created using equal height binning type.
Age (5 bins) and Tenure (2 bins).

```
/* TASK 3, QUESTION 4 */
/* Dummy variables - Age */
data work.fortune_master_dummy; set ana610.fortune_master_clean1;
    if age < 31.0013                        then age_1to31_dum = 1;  else age_1to31_dum = 0;
    if age ge 31.0013 and age < 35.0003 then age_31to35_dum = 1; else age_31to35_dum = 0;
    if age ge 35.0003 and age < 40.0012 then age_35to40_dum = 1; else age_35to40_dum = 0;
    if age ge 40.0012 and age < 47.0016 then age_40to47_dum = 1; else age_40to47_dum = 0;
    if age ge 47.0016                       then age_47plus_dum = 1; else age_47plus_dum = 0;
    if age in(.)                            then age_miss_dum = 1;   else age_miss_dum = 0;
run;
%let age_dums = age_1to31_dum age_31to35_dum age_35to40_dum age_40to47_dum age_47plus_dum age_miss_dum;
proc means data=work.fortune_master_dummy n nmiss min mean max sum; var &age_dums; run;

/* Dummy variables - Tenure */
data work.fortune_master_dummy; set work.fortune_master_dummy;
    if tenure ge 0 and tenure < 7.0014  then tenure_0to7_dum = 1;   else tenure_0to7_dum = 0;
    if tenure ge 7.0014                 then tenure_7plus_dum = 1;  else tenure_7plus_dum = 0;
    if tenure in(.)                     then tenure_miss_dum = 1;   else tenure_miss_dum = 0;
run;
%let tenure_dums = tenure_0to7_dum tenure_7plus_dum tenure_miss_dum;
proc means data=work.fortune_master_dummy n nmiss min mean max sum; var &tenure_dums; run;
data ana610.fortune_master_dummy; set work.fortune_master_dummy;
```

The MEANS Procedure

Variable	N	N Miss	Minimum	Mean	Maximum	Sum
age_1to31_dum	4867	0	0	0.2765564	1.0000000	1346.00
age_31to35_dum	4867	0	0	0.1631395	1.0000000	794.0000000
age_35to40_dum	4867	0	0	0.2046435	1.0000000	996.0000000
age_40to47_dum	4867	0	0	0.1886172	1.0000000	918.0000000
age_47plus_dum	4867	0	0	0.1670434	1.0000000	813.0000000
age_miss_dum	4867	0	0	0.0554757	1.0000000	270.0000000

The MEANS Procedure

Variable	N	N Miss	Minimum	Mean	Maximum	Sum
tenure_0to7_dum	4867	0	0	0.5557839	1.0000000	2705.00
tenure_7plus_dum	4867	0	0	0.4442161	1.0000000	2162.00
tenure_miss_dum	4867	0	0	0	0	0

Transpose - “Tall and skinny”:

```
/* Transpose - Age */
proc transpose data=work.fortune_master_dummy out=work.fortune_master_skinny1; by employee_no;
var &age_dums;
run;
/* Tranpose - Tenure */
proc transpose data=work.fortune_master_dummy out=work.fortune_master_skinny2; by employee_no;
var &tenure_dums;
run;
```

“Tall and skinny” output – Age

CODE LOG RESULTS OUTPUT DATA

Table: WORK.FORTUNE_MASTER_SKINNY1 View: Column names Filter: (none)

Columns: Select all, employee_no, _NAME_, COL1

Property	Value
Label	
Name	
Length	
Type	
Format	
Informat	

Total rows: 29202 Total columns: 3

	employee_no	_NAME_	COL1
1	2316	age_1to31_dum	1
2	2316	age_31to35_dum	0
3	2316	age_35to40_dum	0
4	2316	age_40to47_dum	0
5	2316	age_47plus_dum	0
6	2316	age_miss_dum	0
7	2583	age_1to31_dum	0
8	2583	age_31to35_dum	0
9	2583	age_35to40_dum	0
10	2583	age_40to47_dum	1
11	2583	age_47plus_dum	0
12	2583	age_miss_dum	0
13	2807	age_1to31_dum	0
14	2807	age_31to35_dum	0
15	2807	age_35to40_dum	1
16	2807	age_40to47_dum	0
17	2807	age_47plus_dum	0
18	2807	age_miss_dum	0
19	3361	age_1to31_dum	0
20	3361	age_31to35_dum	0
21	3361	age_35to40_dum	0
22	3361	age_40to47_dum	1
23	3361	age_47plus_dum	0
24	3361	age_miss_dum	0

"Tall and skinny" output – Tenure

Table: WORK.FORTUNE_MASTER_SKINNY2 View: Column names Filter: (none)

Columns: Select all, employee_no, _NAME_, COL1

Property	Value
Label	
Name	
Length	
Type	
Format	

Total rows: 14601 Total columns: 3

	employee_no	_NAME_	COL1
1	2316	tenure_0to7_dum	1
2	2316	tenure_7plus_dum	0
3	2316	tenure_miss_dum	0
4	2583	tenure_0to7_dum	1
5	2583	tenure_7plus_dum	0
6	2583	tenure_miss_dum	0
7	2807	tenure_0to7_dum	1
8	2807	tenure_7plus_dum	0
9	2807	tenure_miss_dum	0
10	3361	tenure_0to7_dum	0
11	3361	tenure_7plus_dum	1
12	3361	tenure_miss_dum	0
13	3408	tenure_0to7_dum	0
14	3408	tenure_7plus_dum	1
15	3408	tenure_miss_dum	0

Check - The frequency of PROC HPBIN matches the sums for each dummy variable bin for Age and Tenure. The left side shows the side-by-side comparison for Age. The right side shows the side-by-side comparison for Tenure.

For the Age variable, during my dummy variable creation, I created a bin for the missing Age values (age_miss_dum) (270 values). PROC HPBIN did not separate a bin for the missing values and it added the 270 missing values to the "age < 31.0013" bin. This explains why "age < 13.0013" in PROC HPBIN has a frequency of 1076, compared to the dummy "age < 13.0013" with a sum of 1346. The 270 value difference is the missing values that PROC HPBIN did not have a separate bin for.

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
ANA610.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	Pseudo-Quantile Binning
Number of Bins Specified	5
Number of Variables	1

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
age	BIN_age	age < 31.0013	1076	0.23406570
		31.0013 <= age < 35.0003	794	0.17272134
		35.0003 <= age < 40.0012	996	0.21666304
		40.0012 <= age < 47.0016	918	0.19969545
		47.0016 <= age	813	0.17685447

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
ANA610.FORTUNE_MASTER_CLEAN1	V9	Input	On Client

Binning Information	
Method	Pseudo-Quantile Binning
Number of Bins Specified	2
Number of Variables	1

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
tenure	BIN_tenure	tenure < 7.0014	2705	0.55578385
		7.0014 <= tenure	2162	0.44421615

The MEANS Procedure

Variable	N	N Miss	Minimum	Mean	Maximum	Sum
age_1to31_dum	4867	0	0	0.2765564	1.0000000	1346.00
age_31to35_dum	4867	0	0	0.1631395	1.0000000	794.0000000
age_35to40_dum	4867	0	0	0.2046435	1.0000000	996.0000000
age_40to47_dum	4867	0	0	0.1886172	1.0000000	918.0000000
age_47plus_dum	4867	0	0	0.1670434	1.0000000	813.0000000
age_miss_dum	4867	0	0	0.0554757	1.0000000	270.0000000

The MEANS Procedure

Variable	N	N Miss	Minimum	Mean	Maximum	Sum
tenure_0to7_dum	4867	0	0	0.5557839	1.0000000	2705.00
tenure_7plus_dum	4867	0	0	0.4442161	1.0000000	2162.00
tenure_miss_dum	4867	0	0	0	0	0

Task #4 (20 pts):

Using PROC CORR and the bin dummy variables your created above for AGE and TENURE:

1. Identify which employee AGE range is most likely to attrition. Support your answer with output from PROC CORR.

   Ages 1-31 are more likely to attrition. The correlation coefficient for the Age 1-31 bin was calculated to be 0.15690, which is the highest in comparison to the other bins. However, 0.15690 is still considered to be a weak, positive correlation because it's far from 1.0.

The CORR Procedure

6 With Variables:	age_1to31_dum age_31to35_dum age_35to40_dum age_40to47_dum age_47plus_dum age_miss_dum
1 Variables:	att_q

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age_1to31_dum	4867	0.27656	0.44734	1346	0	1.00000
age_31to35_dum	4867	0.16314	0.36953	794.00000	0	1.00000
age_35to40_dum	4867	0.20464	0.40348	996.00000	0	1.00000
age_40to47_dum	4867	0.18862	0.39124	918.00000	0	1.00000
age_47plus_dum	4867	0.16704	0.37305	813.00000	0	1.00000
age_miss_dum	4867	0.05548	0.22893	270.00000	0	1.00000
att_q	1470	0.16122	0.36786	237.00000	0	1.00000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	att_q
age_1to31_dum	0.15690 <.0001 1470
age_31to35_dum	0.04803 0.0656 1470
age_35to40_dum	-0.08291 0.0015 1470
age_40to47_dum	-0.09331 0.0003 1470
age_47plus_dum	-0.04860 0.0625 1470
age_miss_dum	0.01604 0.5388 1470

Obs	_NAME_	abs_corr
1	age_1to31_dum	0.15690
2	age_40to47_dum	0.09331
3	age_35to40_dum	0.08291
4	age_47plus_dum	0.04860
5	age_31to35_dum	0.04803
6	age_miss_dum	0.01604

2. Identify which employee TENURE range is most likely to attrition. Support your answer with output from PROC CORR.

Employees with a Tenure of 7+ years are more likely to attrition. The correlation coefficient for the Tenure 7+ bin was calculated to be 0.17807, which is the highest in comparison to the other bins. However, 0.17807 is still considered to be a weak, positive correlation because it's far from 1.0.

The CORR Procedure

2 With Variables:	tenure_1to7_dum tenure_7plus_dum
1 Variables:	att_q

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
tenure_1to7_dum	4867	0.55250	0.49729	2689	0	1.00000
tenure_7plus_dum	4867	0.44422	0.49693	2162	0	1.00000
att_q	1470	0.16122	0.36786	237.00000	0	1.00000

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	att_q
tenure_1to7_dum	0.12762 <.0001 1470
tenure_7plus_dum	-0.17807 <.0001 1470

Obs	_NAME_	abs_corr
1	tenure_7plus_dum	0.17807
2	tenure_1to7_dum	0.12762

Extra Credit (10 pts)

There are 6 obvious extreme values in the variable hire_dt. Can you find them? Explain your logic and provide a table showing employee_number and hire_dt for these 6 "outliers."

Fortune Corp. started its business in June 1980; therefore any hire date before when the company started are obvious outliers. There are 6 "hire_dt" dates that are prior to when the company existed. Employee_no 153573, 211407, 605506, 239944, 601229, and 51481 are the 6 outliers.

Total rows: 4867 Total columns: 2

	employee_no	hire_dt
1	153573	10/10/75
2	211407	08/03/76
3	605506	10/06/79
4	239944	10/27/79
5	601229	12/07/79
6	51481	02/10/80
7	475097	06/05/80
8	53364	07/04/80

Homework deliverables:

Task #1

- Neatly formatted, data audit report (see template) – Word doc
- Merged SAS datafile (download from SAS Studio)

Task #2

- An additional Word doc with your analysis and discussion, including all tables and charts
- A SAS program with all code used for this assignment (both Task 1 and 2)