Data Audit Report

Prepared by

Magnus Aghe

In support of

Predictive Model to Test Employee Voluntary Attrition

Requested by

SVP of Human Resources

Fortune Corporation

March 28, 2023

Introduction

The analytics team has been asked by the SVP of Human Resources at Fortune Corp, maker of specialized laboratory equipment for the pharmaceutical industry, to build a predictive model and test for employee voluntary attrition.

The target sample qualifications provided by the SVP of Human Resources, are those employees that have 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

From this sample, the target segments for modeling will be:

- Yes/event (1): yes, voluntarily attritioned
- No/non-event (0): no, did not attrition

Founded in 1980, Fortune Corp prides itself on employee job satisfaction and now seeks to understand why employees voluntarily leave the company. To aid this objective, the analytics team has been provided with several datafiles by the IT department.

The supplied datafiles are intended to support the development of a model that would help find/score current employees who might be thinking of leaving, so that proactive steps can be taken to retain them.

The purpose of this data audit is to ensure the following:

1. The analytics team has received all datafiles intended for this project.
2. The analytics team understands the content, layout, and format of these files.
3. The data in these files are of sufficient integrity and quantity to support the model development.

This data audit consists of 4 sections:

1. **Datafile Summary**: A list and description of all datafiles received.
2. **Datafile Detail**: For each datafile, tables showing all data fields received, their values, summary statistics, and distributions. Data fields are categorized into one of 4 types of analytical variables:
   - Categorical - data fields with distinct levels or values which represent categories; can be a number or a label, nominal or ordinal.
   - Date - data fields that are identified as calendar dates.
   - Numeric - data fields that are continuous numeric data.
   - Character - data fields whose values are characters and are not otherwise classified as categorical.
3. **Modeling Sample** – After merging all supplied datafiles, a determination is made as to whether there is adequate sample size for each target sample to support model development.
4. **Questions** – The auditing process will uncover data integrity issues. This section lists what the analytics team has found in this regard. This section also poses specific questions on data field definitions, field coding, and interpretation, answers to which will facilitate the team's model development effort.

Datafile Summary

The analytics team has received 5 datafiles from Fortune Corp IT department as listed in Table 1.

Table 1. Datafiles Received

Filename	File Type	# Of Records	File Contents
fortune_credit	CSV	4, 867	FICO Score, SSN
fortune_acct	SAS	4, 867	Employee Number, SSN, Department, Monthly Income, other misc. Account Vars
fortune_attrition	SAS	262	Employee Number, Departure Date
fortune_hr	SAS	4, 867	Employee Number, First Name, Gender, Hire Date, other misc. Employee Vars
fortune_survey	SAS	1, 470	Employee Number, Job Level, Total Working Years, other Survey Vars

Datafile Detail

Each datafile contains the analytic data fields as shown in the following tables. Note that the data fields have been classified based on their potential analytical usage.

Datafile #1: **Credit Bureau file (fortune_credit)**

File Analytic Contents:

Numeric Fields (2): FICO_SCR, SSN

Categorical Fields (0):

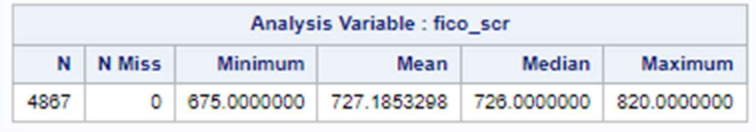
Character Fields (0):

Date Fields (0):

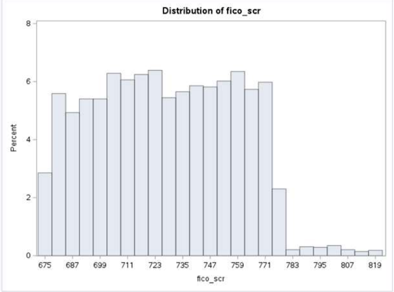
Records: 4,867
Columns: 2
Notes: The data field SSN appears to be a row id or index field.

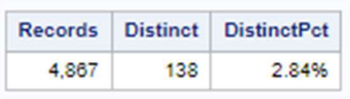
Table 2. fortune_credit - Numeric Data

Analytic numeric data fields

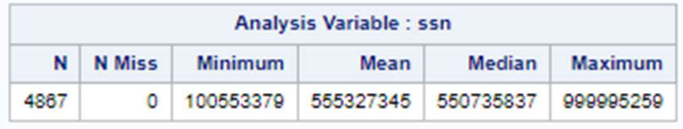


Analysis Variable : fico_scr					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	675.0000000	727.1853298	726.0000000	820.0000000

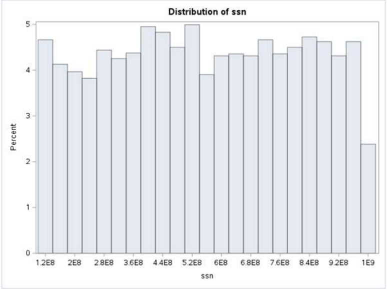


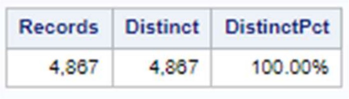


Records	Distinct	DistinctPct
4,867	138	2.84%



Analysis Variable : ssn					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	100553379	555327345	550735837	999995259





Records	Distinct	DistinctPct
4,867	4,867	100.00%

***Datafile #2:* Accounting file (fortune_acct)**

File Analytic Contents:

Numeric Fields (5): DAILYRATE, HOURLYRATE, MONTHLYINCOME, PERCENTSALARYHIKE, EMPLOYEE_NO

Categorical Fields (4): PERFORMANCERATING, STOCKOPTIONLEVEL, DEPARTMENT, OVERTIME

Character Fields (1): SSN

Date Fields (0):

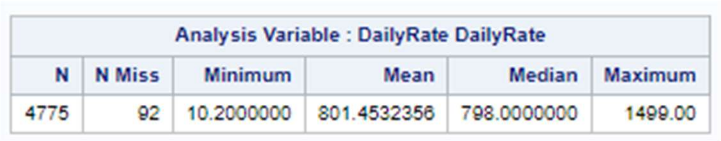
Records: 4,867

Columns: 10

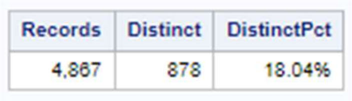
Notes: The data field EMPLOYEE_NO appears to be a row id or index field.

Table 3. fortune_acct - Numeric Data

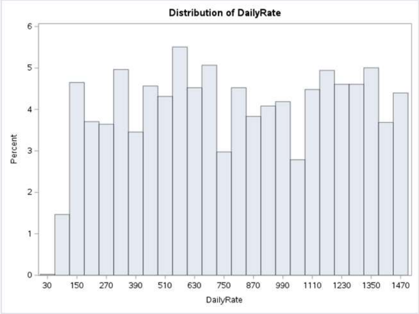
Analytic numeric data fields

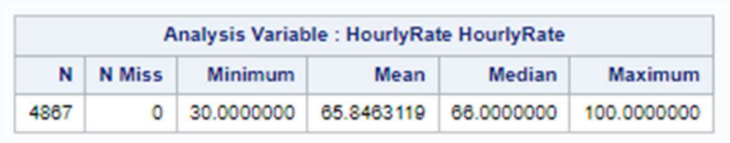


Analysis Variable : DailyRate DailyRate					
N	N Miss	Minimum	Mean	Median	Maximum
4775	92	10.2000000	801.4532356	798.0000000	1499.00

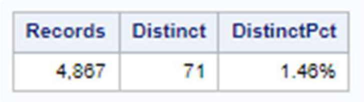


Records	Distinct	DistinctPct
4,867	878	18.04%

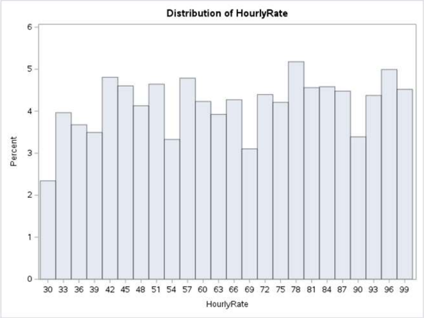


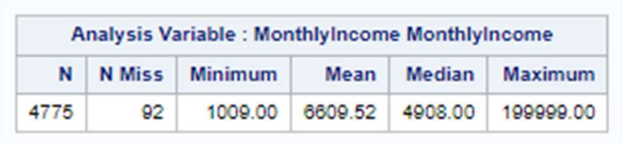


Analysis Variable : HourlyRate HourlyRate					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	30.0000000	65.8463119	66.0000000	100.0000000

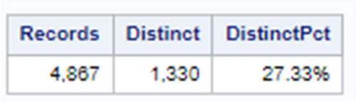


Records	Distinct	DistinctPct
4,867	71	1.46%

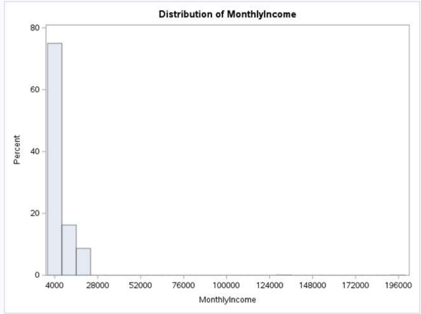


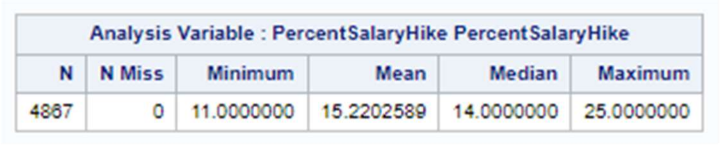


Analysis Variable : MonthlyIncome MonthlyIncome					
N	N Miss	Minimum	Mean	Median	Maximum
4775	92	1009.00	6609.52	4908.00	199999.00

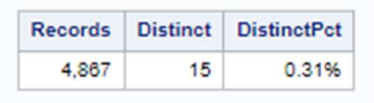


Records	Distinct	DistinctPct
4,867	1,330	27.33%

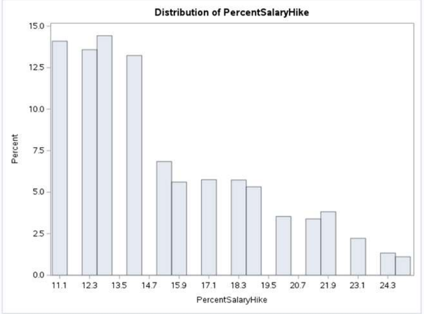


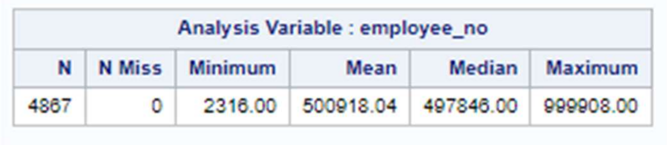


Analysis Variable : PercentSalaryHike PercentSalaryHike					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	11.0000000	15.2202589	14.0000000	25.0000000

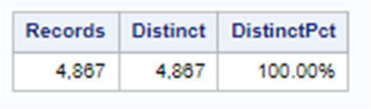


Records	Distinct	DistinctPct
4,867	15	0.31%





Analysis Variable : employee_no					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	2316.00	500918.04	497846.00	999908.00



Records	Distinct	DistinctPct
4,867	4,867	100.00%

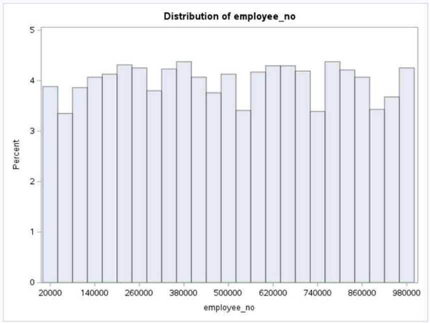
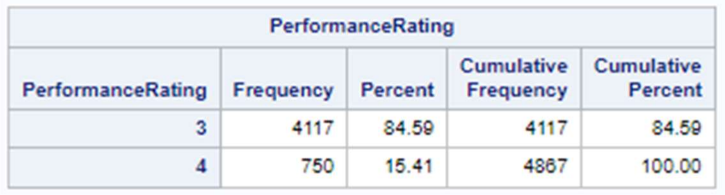
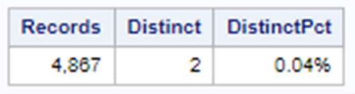


Table 4. fortune_acct - Categorical Data

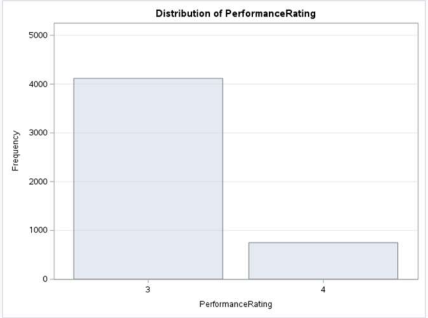
Analytic categorical data fields

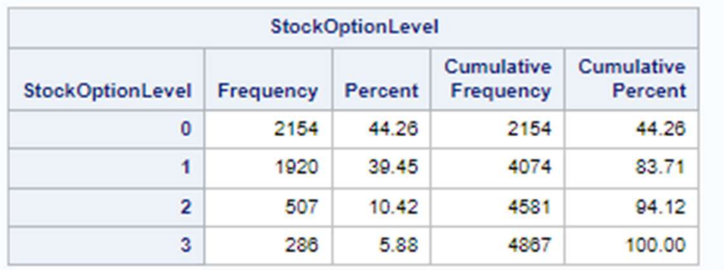


PerformanceRating				
PerformanceRating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	4117	84.59	4117	84.59
4	750	15.41	4867	100.00

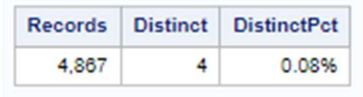


Records	Distinct	DistinctPct
4,867	2	0.04%

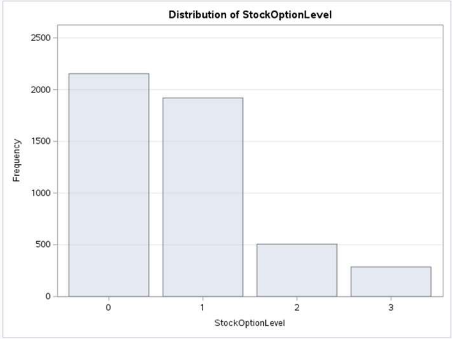


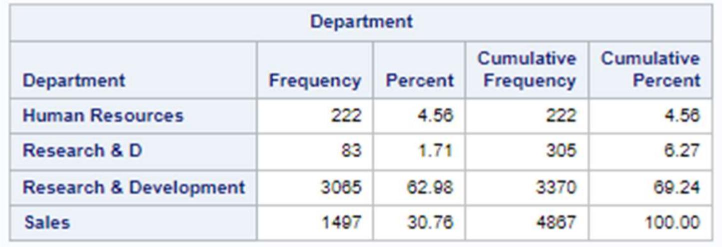


StockOptionLevel				
StockOptionLevel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2154	44.26	2154	44.26
1	1920	39.45	4074	83.71
2	507	10.42	4581	94.12
3	286	5.88	4867	100.00

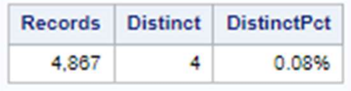


Records	Distinct	DistinctPct
4,867	4	0.08%

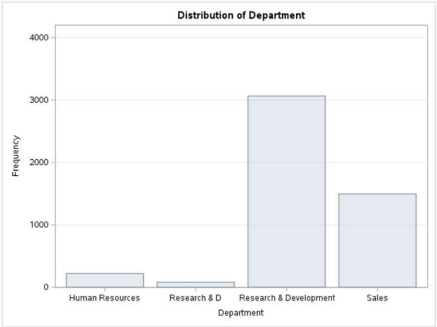


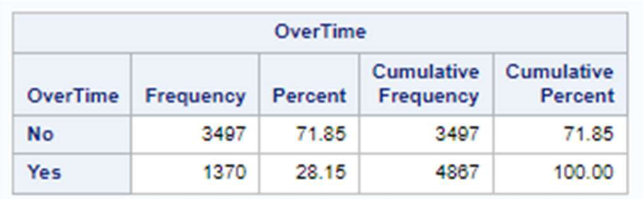


Department				
Department	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Human Resources	222	4.56	222	4.56
Research & D	83	1.71	305	6.27
Research & Development	3065	62.98	3370	69.24
Sales	1497	30.76	4867	100.00

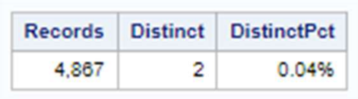


Records	Distinct	DistinctPct
4,867	4	0.08%





OverTime				
OverTime	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	3497	71.85	3497	71.85
Yes	1370	28.15	4867	100.00



Records	Distinct	DistinctPct
4,867	2	0.04%

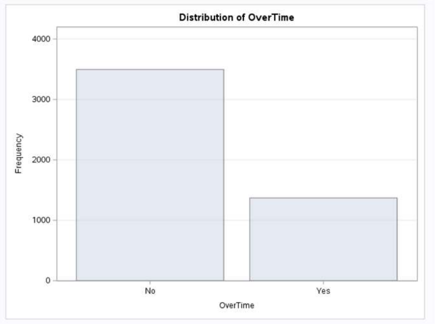
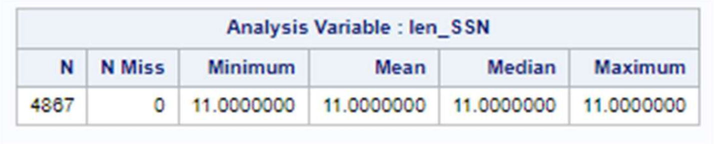
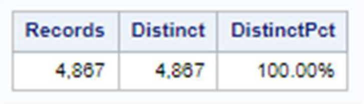


Table 5. fortune_acct - Character Data

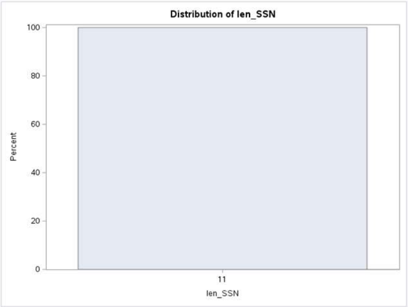
Analytic character data fields



Analysis Variable : len_SSN					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	11.0000000	11.0000000	11.0000000	11.0000000



Records	Distinct	DistinctPct
4,867	4,867	100.00%



Datafile #3: **Attrition file (fortune_attrition)**

File Analytic Contents:

Numeric Fields (1): EMPLOYEE_NO

Categorical Fields (0):

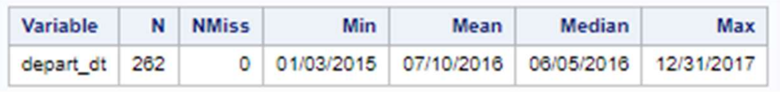
Character Fields (0):

Date Fields (1): DEPART_DT

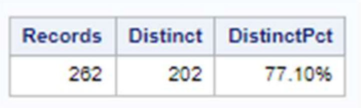
Records: 262
Columns: 2
Notes: The data field EMPLOYEE_NO appears to be a row id or index field.

Table 6. fortune_attrition - Date Data

Analytic date data fields



Variable	N	NMiss	Min	Mean	Median	Max
depart_dt	262	0	01/03/2015	07/10/2016	06/05/2016	12/31/2017



Records	Distinct	DistinctPct
262	202	77.10%

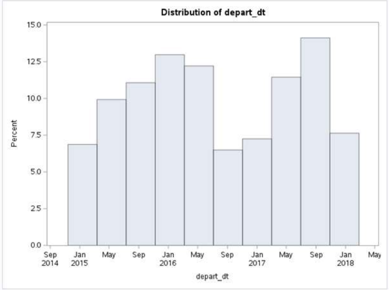
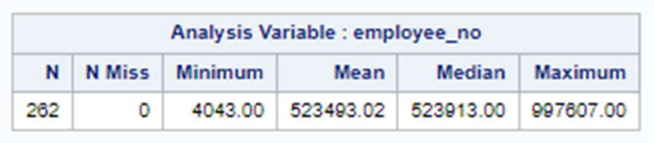


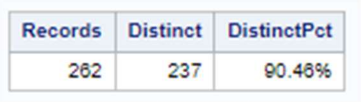
Table 7. fortune_attrition - Numeric Data

Analytic numeric data fields

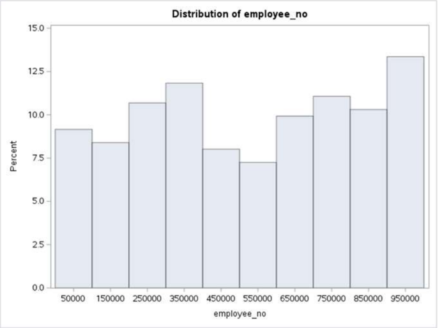


Analysis Variable : employee_no

N	N Miss	Minimum	Mean	Median	Maximum
262	0	4043.00	523493.02	523913.00	997607.00



Records	Distinct	DistinctPct
262	237	90.46%



Datafile #4: HR file (fortune_hr)

File Analytic Contents:

Numeric Fields (1): EMPLOYEE_NO

Categorical Fields (4): EDUCATION, EDUCATIONFIELD, GENDER, BIRTH_STATE

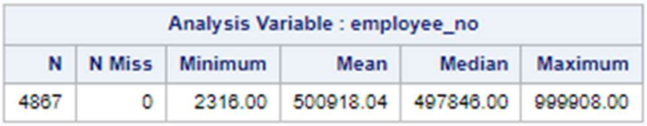
Character Fields (1): FIRST_NAME

Date Fields (2): BIRTH_DT, HIRE_DT

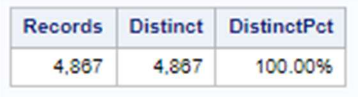
Records: 4,867
Columns: 8
Notes: The data field EMPLOYEE_NO appears to be a row id or index field.

Table 8. fortune_hr - Numeric Data

Analytic numeric data fields



Analysis Variable : employee_no					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	2316.00	500918.04	497846.00	999908.00



Records	Distinct	DistinctPct
4,867	4,867	100.00%

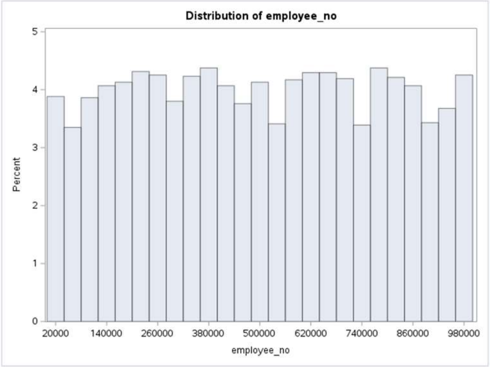
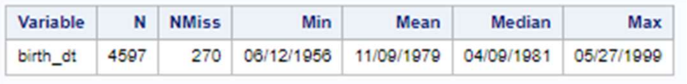
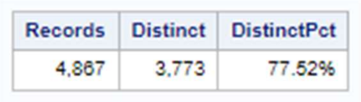


Table 9. fortune_hr - Date Data

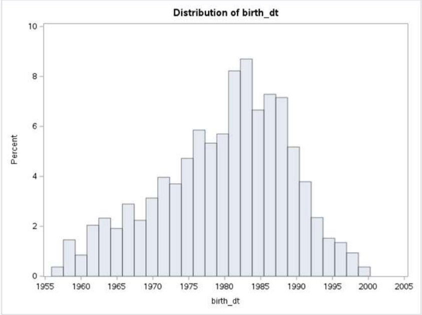
Analytic date data fields

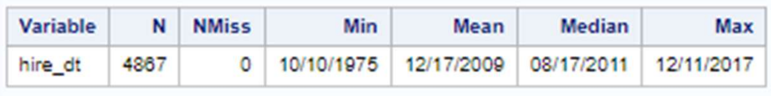


Variable	N	NMiss	Min	Mean	Median	Max
birth_dt	4597	270	06/12/1956	11/09/1979	04/09/1981	05/27/1999

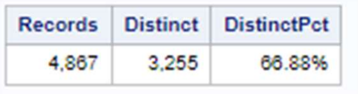


Records	Distinct	DistinctPct
4,867	3,773	77.52%





Variable	N	NMiss	Min	Mean	Median	Max
hire_dt	4867	0	10/10/1975	12/17/2009	08/17/2011	12/11/2017



Records	Distinct	DistinctPct
4,867	3,255	66.88%

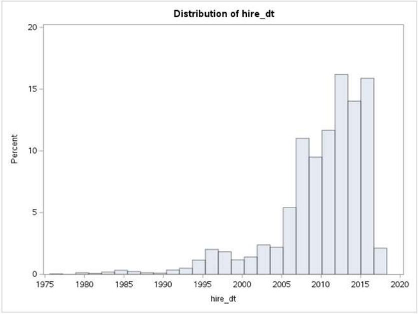
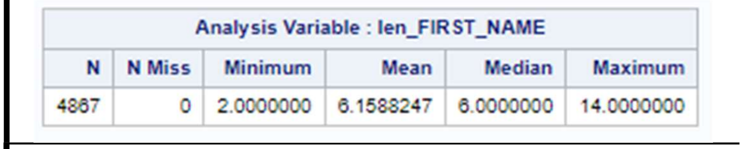
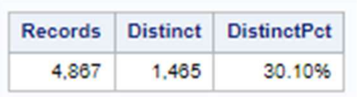


Table 10. fortune_hr - Character Data

Analytic character data fields



Analysis Variable : len_FIRST_NAME					
N	N Miss	Minimum	Mean	Median	Maximum
4867	0	2.0000000	6.1588247	6.0000000	14.0000000



Records	Distinct	DistinctPct
4,867	1,465	30.10%

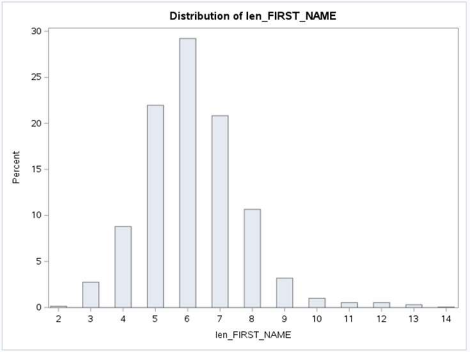
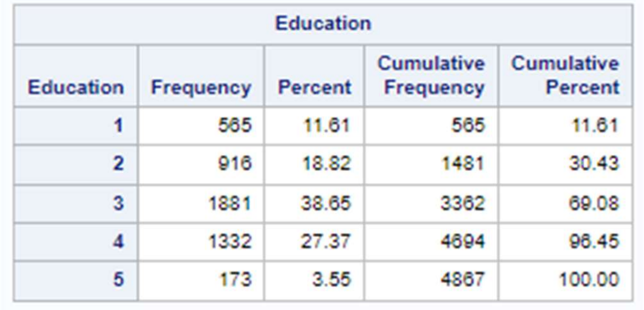
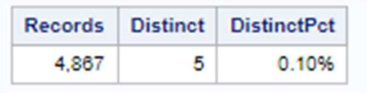


Table 11. fortune_hr - Categorical Data

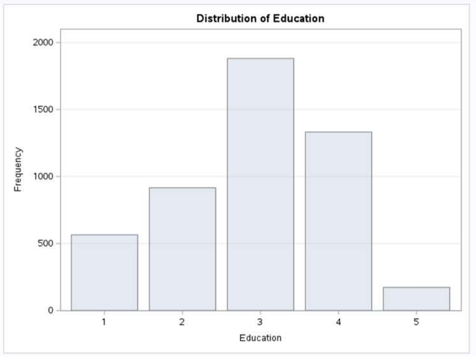
Analytic categorical data fields

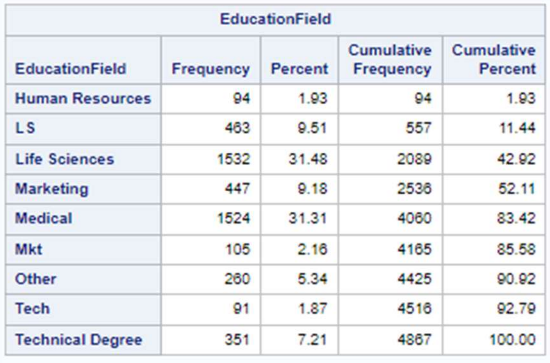


Education				
Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	565	11.61	565	11.61
2	916	18.82	1481	30.43
3	1881	38.65	3362	69.08
4	1332	27.37	4694	96.45
5	173	3.55	4867	100.00

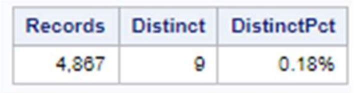


Records	Distinct	DistinctPct
4,867	5	0.10%

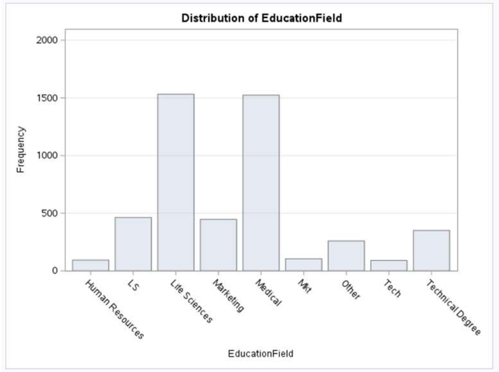


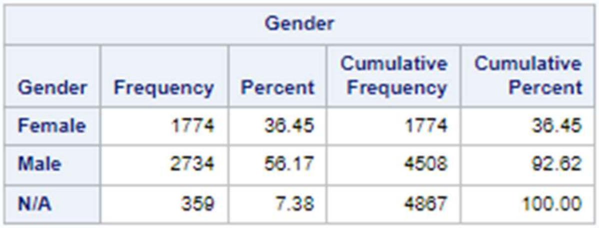


EducationField				
EducationField	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Human Resources	94	1.93	94	1.93
LS	463	9.51	557	11.44
Life Sciences	1532	31.48	2089	42.92
Marketing	447	9.18	2536	52.11
Medical	1524	31.31	4060	83.42
Mkt	105	2.16	4165	85.58
Other	260	5.34	4425	90.92
Tech	91	1.87	4516	92.79
Technical Degree	351	7.21	4867	100.00

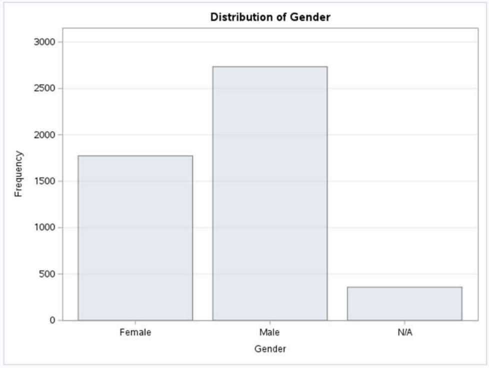


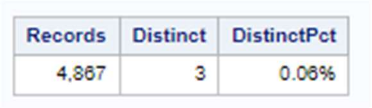
Records	Distinct	DistinctPct
4,867	9	0.18%



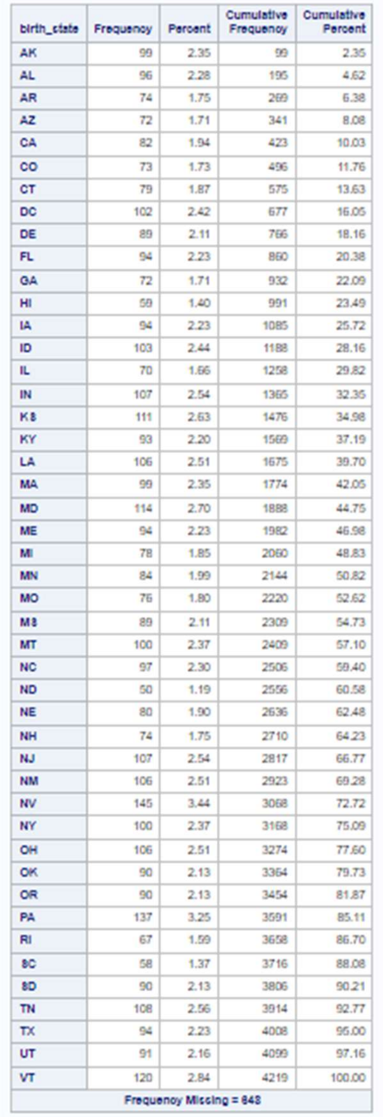


Gender				
Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	1774	36.45	1774	36.45
Male	2734	56.17	4508	92.62
N/A	359	7.38	4867	100.00

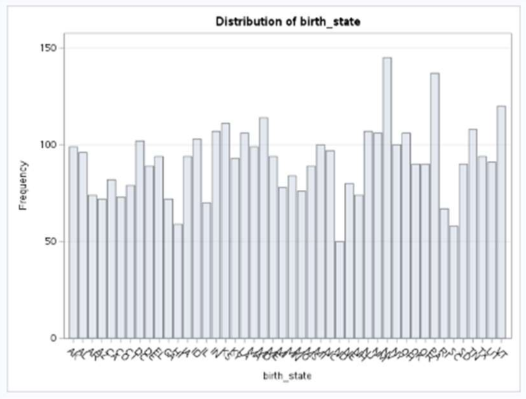


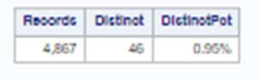


Records	Distinct	DistinctPct
4,867	3	0.06%



birth_state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AK	99	2.35	99	2.35
AL	96	2.28	195	4.62
AR	74	1.75	269	6.38
AZ	72	1.71	341	8.08
CA	82	1.94	423	10.03
CO	73	1.73	496	11.76
CT	79	1.87	575	13.63
DC	102	2.42	677	16.05
DE	89	2.11	766	18.16
FL	94	2.23	860	20.38
GA	72	1.71	932	22.09
HI	59	1.40	991	23.49
IA	94	2.23	1085	25.72
ID	103	2.44	1188	28.16
IL	70	1.66	1258	29.82
IN	107	2.54	1365	32.35
KS	111	2.63	1476	34.98
KY	93	2.20	1569	37.19
LA	106	2.51	1675	39.70
MA	99	2.35	1774	42.05
MD	114	2.70	1888	44.75
ME	94	2.23	1982	46.98
MI	78	1.85	2060	48.83
MN	84	1.99	2144	50.82
MO	76	1.80	2220	52.62
MS	89	2.11	2309	54.73
MT	100	2.37	2409	57.10
NC	97	2.30	2506	59.40
ND	50	1.19	2556	60.58
NE	80	1.90	2636	62.48
NH	74	1.75	2710	64.23
NJ	107	2.54	2817	66.77
NM	106	2.51	2923	69.28
NV	145	3.44	3068	72.72
NY	100	2.37	3168	75.09
OH	106	2.51	3274	77.60
OK	90	2.13	3364	79.73
OR	90	2.13	3454	81.87
PA	137	3.25	3591	85.11
RI	67	1.59	3658	86.70
SC	58	1.37	3716	88.08
SD	90	2.13	3806	90.21
TN	108	2.56	3914	92.77
TX	94	2.23	4008	95.00
UT	91	2.16	4099	97.16
VT	120	2.84	4219	100.00
Frequency Missing = 648				





Records	Distinct	DistinctPct
4,867	46	0.95%

Datafile #5: **Survey file (fortune_survey)**

File Analytic Contents:

Numeric Fields (8): DISTANCEFROMHOME, NUMCOMPANIESWORKED, TOTALWORKINGYEARS, TRAININGTIMESLASTYEAR, YEARSINCURRENTROLE, YEARSSINCELASTPROMOTION, YEARSWITHCURRMANAGER, EMPLOYEE_NO

Categorical Fields (8): BUSINESSTRAVEL, ENVIRONMENTSATISFACTION, JOBINVOLVEMENT, JOBLEVEL, JOBSATISFACTION, MARITALSTATUS, RELATIONSHIPSATISFACTION, WORKLIFEBALANCE

Character Fields (0):

Date Fields (0):

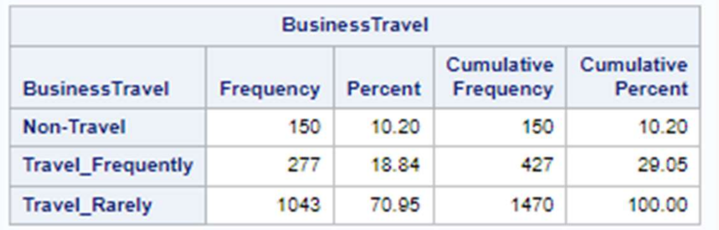
Records: 1,470

Columns: 16

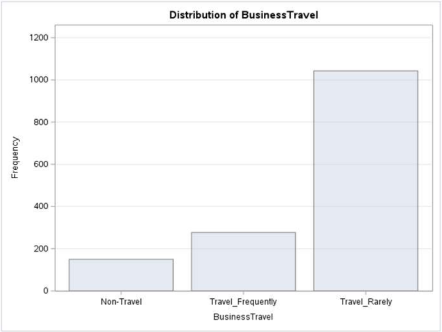
Notes: The data field EMPLOYEE_NO appears to be a row id or index field.

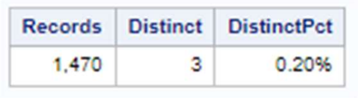
Table 12. fortune_survey - Categorical Data

Analytic categorical data fields

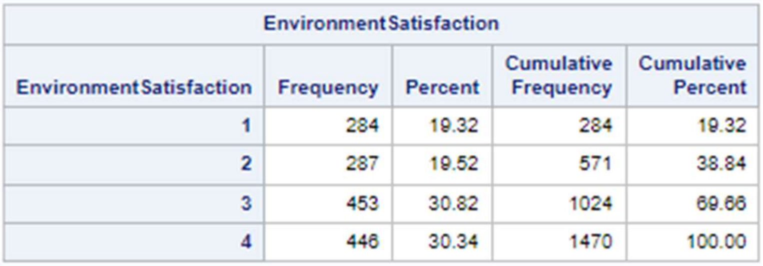


BusinessTravel				
BusinessTravel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Non-Travel	150	10.20	150	10.20
Travel_Frequently	277	18.84	427	29.05
Travel_Rarely	1043	70.95	1470	100.00

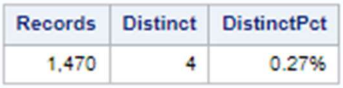




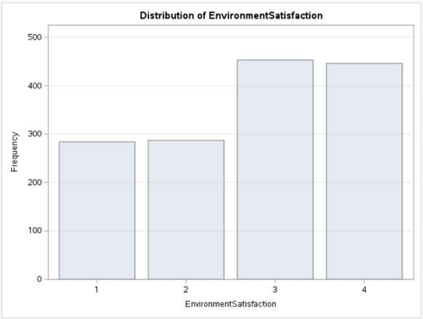
Records	Distinct	DistinctPct
1,470	3	0.20%

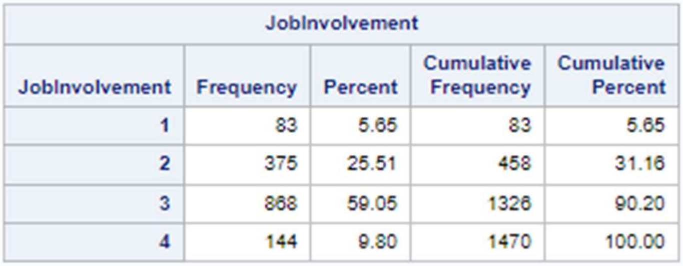


EnvironmentSatisfaction				
EnvironmentSatisfaction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	284	19.32	284	19.32
2	287	19.52	571	38.84
3	453	30.82	1024	69.66
4	446	30.34	1470	100.00

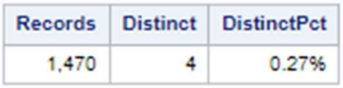


Records	Distinct	DistinctPct
1,470	4	0.27%

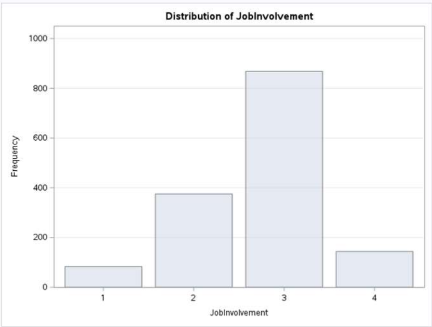


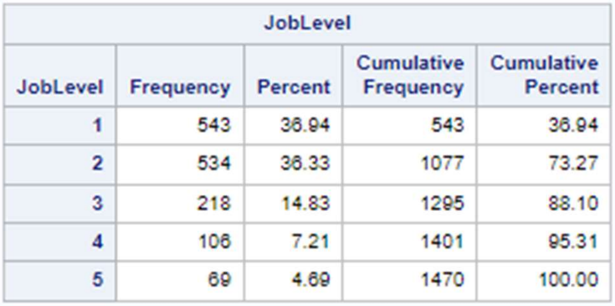


JobInvolvement				
JobInvolvement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	83	5.65	83	5.65
2	375	25.51	458	31.16
3	868	59.05	1326	90.20
4	144	9.80	1470	100.00

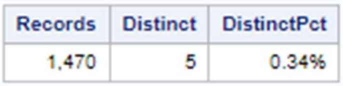


Records	Distinct	DistinctPct
1,470	4	0.27%

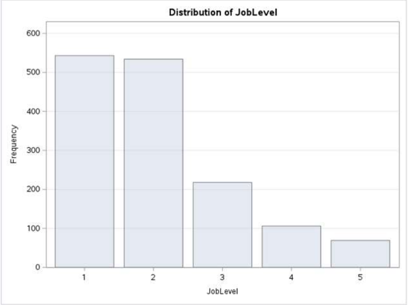


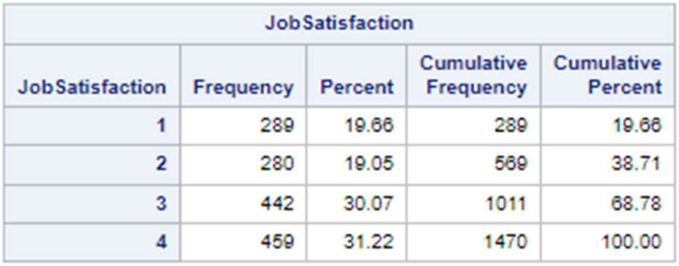


JobLevel				
JobLevel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	543	36.94	543	36.94
2	534	36.33	1077	73.27
3	218	14.83	1295	88.10
4	106	7.21	1401	95.31
5	69	4.69	1470	100.00

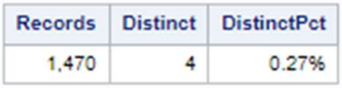


Records	Distinct	DistinctPct
1,470	5	0.34%

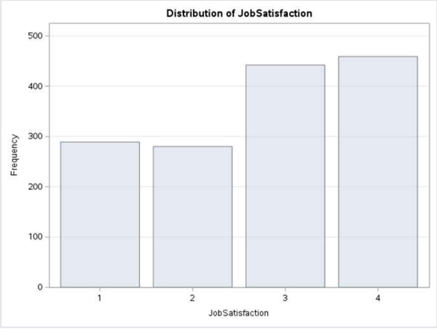


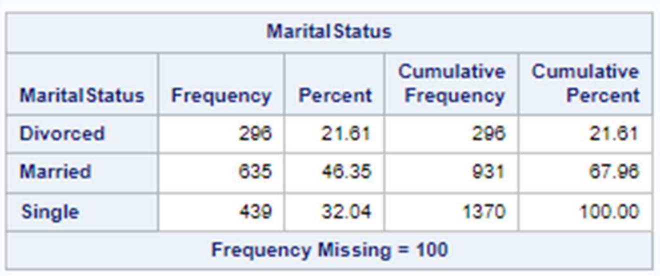


JobSatisfaction				
JobSatisfaction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	289	19.66	289	19.66
2	280	19.05	569	38.71
3	442	30.07	1011	68.78
4	459	31.22	1470	100.00

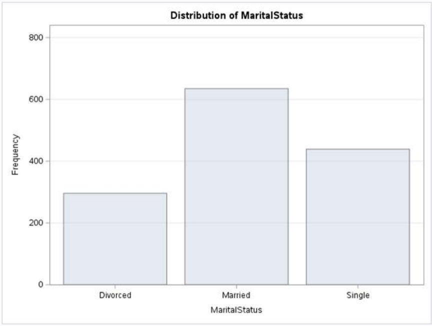


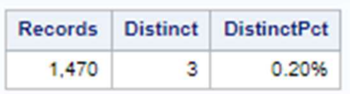
Records	Distinct	DistinctPct
1,470	4	0.27%



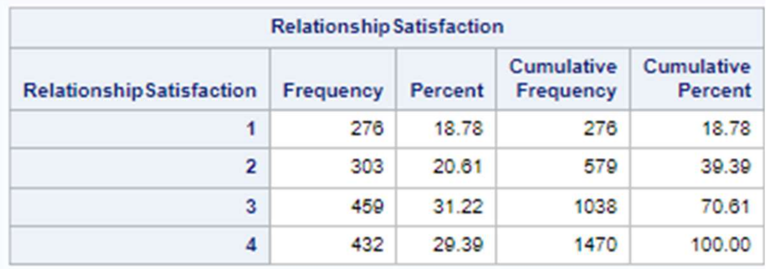


MaritalStatus				
MaritalStatus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	296	21.61	296	21.61
Married	635	46.35	931	67.96
Single	439	32.04	1370	100.00
Frequency Missing = 100				

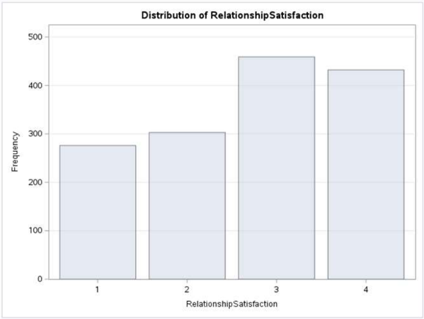


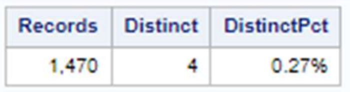


Records	Distinct	DistinctPct
1,470	3	0.20%

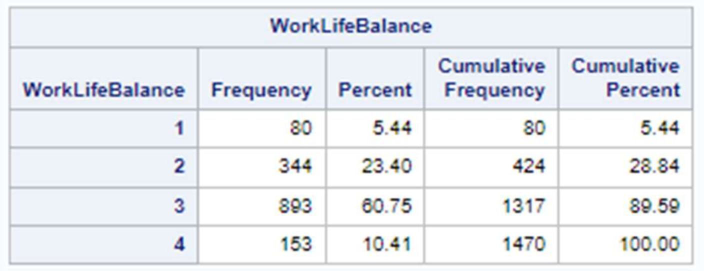


RelationshipSatisfaction				
RelationshipSatisfaction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	276	18.78	276	18.78
2	303	20.61	579	39.39
3	459	31.22	1038	70.61
4	432	29.39	1470	100.00

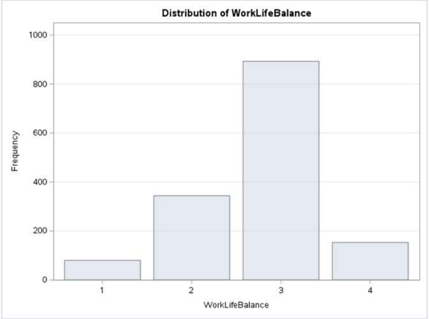


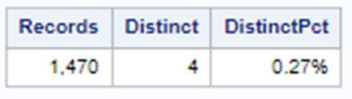


Records	Distinct	DistinctPct
1,470	4	0.27%



WorkLifeBalance				
WorkLifeBalance	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	80	5.44	80	5.44
2	344	23.40	424	28.84
3	893	60.75	1317	89.59
4	153	10.41	1470	100.00

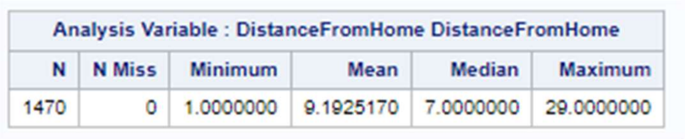




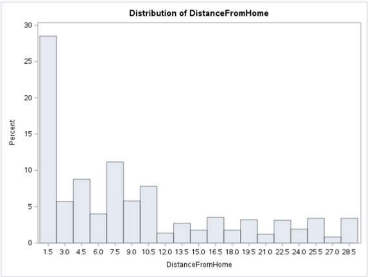
Records	Distinct	DistinctPct
1,470	4	0.27%

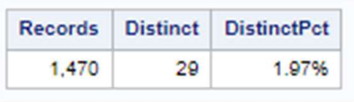
Table 13. fortune_survey - Numeric Data

Analytic numericl data fields

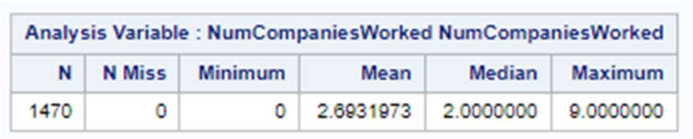


Analysis Variable : DistanceFromHome DistanceFromHome					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	1.0000000	9.1925170	7.0000000	29.0000000

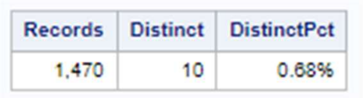




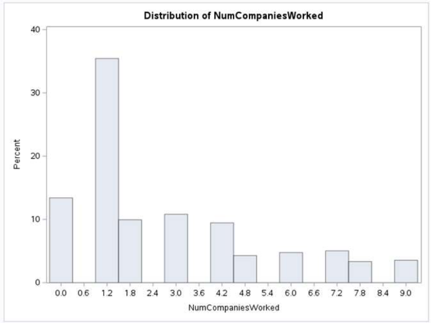
Records	Distinct	DistinctPct
1,470	29	1.97%

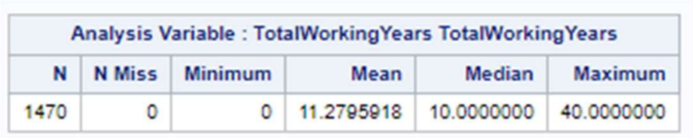


Analysis Variable : NumCompaniesWorked NumCompaniesWorked					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	2.6931973	2.0000000	9.0000000

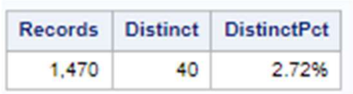


Records	Distinct	DistinctPct
1,470	10	0.68%

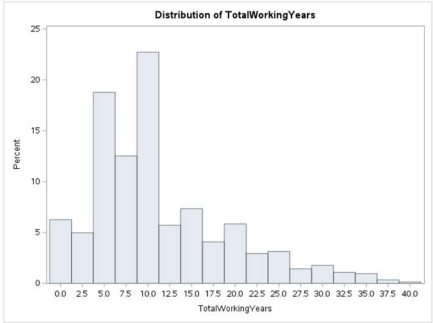


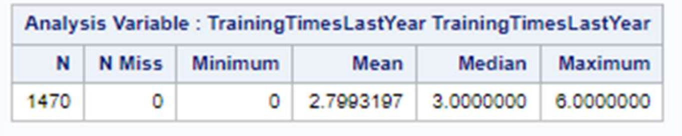


Analysis Variable : TotalWorkingYears TotalWorkingYears					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	11.2795918	10.0000000	40.0000000

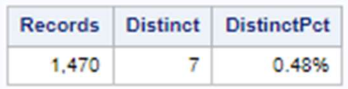


Records	Distinct	DistinctPct
1,470	40	2.72%

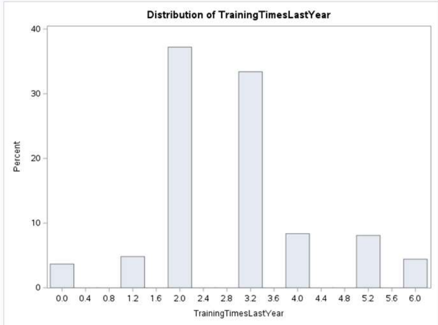


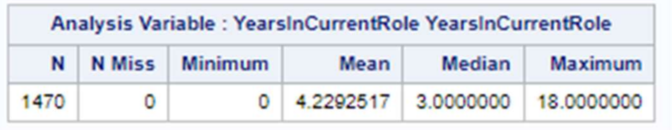


Analysis Variable : TrainingTimesLastYear TrainingTimesLastYear					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	2.7993197	3.0000000	6.0000000

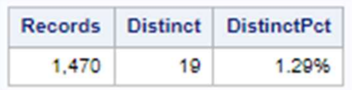


Records	Distinct	DistinctPct
1,470	7	0.48%

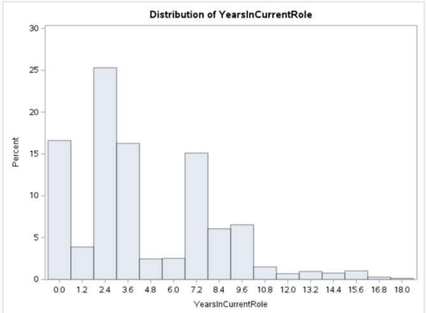


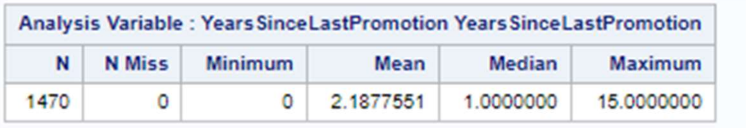


Analysis Variable : YearsInCurrentRole YearsInCurrentRole					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	4.2292517	3.0000000	18.0000000

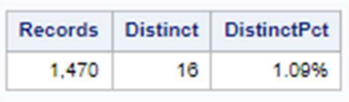


Records	Distinct	DistinctPct
1,470	19	1.29%

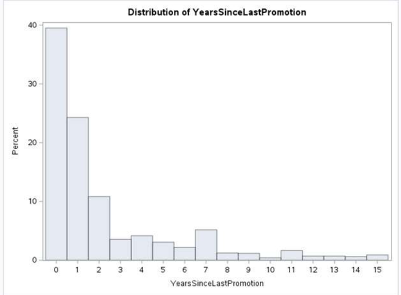


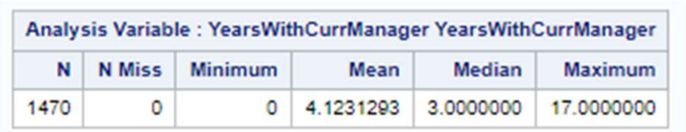


Analysis Variable : YearsSinceLastPromotion YearsSinceLastPromotion					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	2.1877551	1.0000000	15.0000000

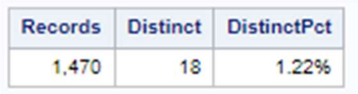


Records	Distinct	DistinctPct
1,470	16	1.09%

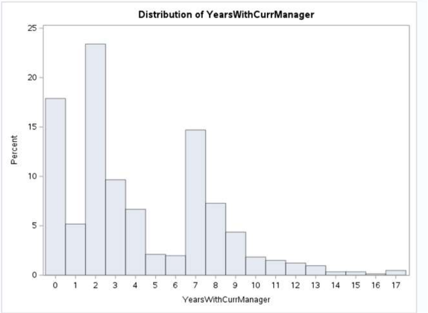


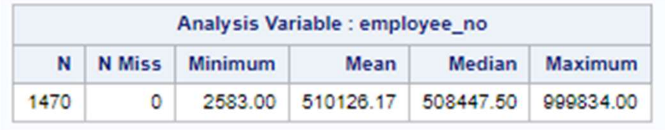


Analysis Variable : YearsWithCurrManager YearsWithCurrManager					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	0	4.1231293	3.0000000	17.0000000

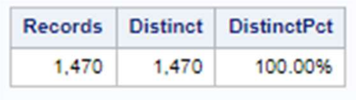


Records	Distinct	DistinctPct
1,470	18	1.22%

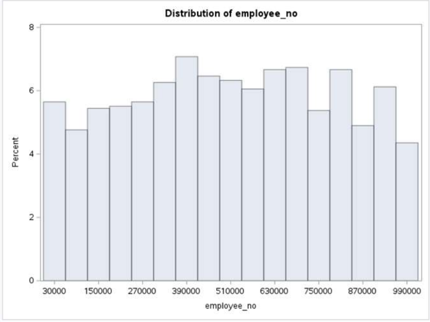




Analysis Variable : employee_no					
N	N Miss	Minimum	Mean	Median	Maximum
1470	0	2583.00	510126.17	508447.50	999834.00



Records	Distinct	DistinctPct
1,470	1,470	100.00%



Modeling Sample

Segment	Count
Available event (yes) sample	262
Available non-event (no) sample	1, 233
Total (target) sample	1, 495
Total records in dataset	4, 892

Sample size in the event target segment is not adequate to support the predictive model since it has less than 1,000 observations even though the non-event target segment has over 1,000 observations.

Questions

1. Does the above information appear to be correct? Specifically:
   - o Does the analytics team have all the data that was meant to be sent?
   - o Is the team interpreting the data correctly?
   - o Do the data appear to have reasonable values?

2. Here is a list of the data integrity issues the analytics team uncovered. Please review:

   a. DailyRate (fortune_acct file)
      - o Missing Values – 92 (1.89% of the dataset)
      - o Extreme Values (Low) – Minimum of 10.2 (where mean is 801.45)?
      - o Extreme Values (High) – Maximum of 1,499 (where mean is 801.45)

   b. MonthlyIncome (fortune_acct file)
      - o Missing Values – 92 (1.89% of the dataset)
      - o Extreme Values (High) – Maximum of 199,999 a month as income?

   c. Department (fortune_acct file)
      - o Extreme Values (Low) – Only 83 employees in Research & D department i.e., 1.71% of dataset? Perhaps there's an error in SEPARATING "Research & D" from "Research & Development"?

   d. birth_dt (fortune_hr file)
      - o Missing Values – 270 (5.55% of the dataset)

   e. hire_dt (fortune_hr file)
      - o Extreme Values (Low) – Earliest hire date is 10/10/1975. How is this possible when the company opened for business in June 1980?

   f. Gender (fortune_hr file)
      - o Odd Values – N/A had a frequency of 359 (7.18% of dataset). What does N/A mean?

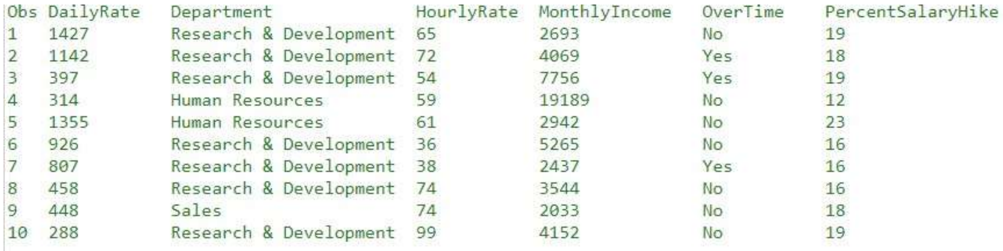
   g. birth_state (fortune_hr file)
      - o Missing values – 648 (13.3% of the dataset)

   h. MaritalStatus (fortune_survey file)
      - o Missing Values – 100 (out of 1470) i.e., 6.8% of dataset.

3. The following are specific questions the analytics team has about the data. Please review:

   a. DailyRate (fortune_acct file)
      - Does this denote wages per day, or does it measure some other metric? If wages/salary, on what basis is DailyRate being computed, is it 8 hours per day or as a fraction of monthly income? Because there seem to be no correlation of daily rate with hourly rate or monthly income? Or are decimal places missing in the data?

      - Observation 1 below highlights this (DailyRate = 1427?, HourlyRate = 65, MonthlyIncome = 2693). It would take working almost 22 hours at a rate of $65 per hour to earn $1,427 in a day! Similarly, a DailyRate of 1,427 would earn more than the observed MonthlyIncome in 2 days!



```
Obs DailyRate  Department              HourlyRate MonthlyIncome OverTime  PercentSalaryHike
1   1427       Research & Development  65         2693          No        19
2   1142       Research & Development  72         4069          Yes       18
3   397        Research & Development  54         7756          Yes       19
4   314        Human Resources         59         19189         No        12
5   1355       Human Resources         61         2942          No        23
6   926        Research & Development  36         5265          No        16
7   807        Research & Development  38         2437          Yes       16
8   458        Research & Development  74         3544          No        16
9   448        Sales                   74         2033          No        18
10  288        Research & Development  99         4152          No        19
```

   b. PerformanceRating (fortune_acct file)
      - What does 3 and 4 represent? They are the only distinct values in the data.

   c. StockOptionLevel (fortune_acct file)
      - What does 0, 1, 2, 3 represent? What is the order of ranking?

   d. Department (fortune_acct file)
      - Are there 3 or 4 departments? Is “Research & D” not the same as “Research & Development”? Should we merge Research & D into Research & Development, so that we have 3 departments instead of 4.

   e. employee_no (fortune_attrition file)
      - There are 237 distinct employees but 262 who attritioned. Does this mean that there were some employees who got hired more than once and also attritioned (left the company) more than once? Or is this some error?

   f. Education (fortune_hr file)
      - What do the values 1, 2, 3, 4, 5 represent?

g. EnvironmentSatisfaction (fortune_survey file)
   - What does the ranking from 1 to 4 imply?

h. JobInvolvement (fortune_survey file)
   - What does the ranking from 1 to 4 imply?

i. JobLevel (fortune_survey file)
   - What does the ranking from 1 to 5 represent?

j. JobSatisfaction (fortune_survey file)
   - What does the ranking from 1 to 4 imply?

k. RelationshipSatisfaction (fortune_survey file)
   - What does the ranking from 1 to 4 imply?

l. WorkLifeBalance (fortune_survey file)
   - What does the ranking from 1 to 4 represent?