Data Management and Visualization Exam

Morteza Bakhtiari

July 13, 2025

Contents

List of Tables

List of Figures

1 Exercise I: Document Setup

This section shows the setup for the content of the exam problem set

- A table of contents, list of tables, and list of figures

- Proper section organization
- Descriptive text explaining the setup process

The setup required several LaTeX packages:

- `fancyhdr` for custom headers and footers
- `lastpage` for page numbering
- `graphicx` for figure inclusion
- `hyperref` for better PDF navigation

2 Exercise II: Data Preparation and Cleaning

2.1 Data Summary

This section contain the data preparation with tables and figures. This dataset includes 1044950 observations of 83 variables.
Key Variables:

- **Demographics:**
  - personid: Unique identifier (character)
  - year: Ranges from 1993 to 2013 (median = 2003)
  - age: 23–99 years (mean = 45.1, but has 52,277 NA values)
  - gender, raceth, bthus (Place of birth): Categorical
- **Education:**
  - dgrdg (Type of highest certificate or degree)
  - hd03y5 (Year of highest degree 2003-onward (5 year intervals))
  - ndgmed (Field of major for highest degree): Categorical
- **Employment:**
  - hrswkgr: (hours per week typically worked (group)).
  - salary: Numeric (salary has extreme values: median =

  - jobsatis, satsal (job/salary satisfaction): Categorical

- **Job Characteristics**
  - wamgmt (Work activities on principal job: management and administration)
  - waprod (Work activities on principal job: production, operations, maintenance)
  - watea (Work activities on principal job: teaching): Categorical.

- **Notable Observations:**
  - Missing Data:
    * age has 52,277 NAs
    * salary has 52,297 NAs
  - Time Coverage: Data spans 1993–2013 (likely longitudinal/panel data).
  - Data class: Categorical Dominance: Most variables are character type (may need conversion to factors for analysis).

Table 1: Top 10 Missisng summary data

	variable	missing_count	missing_pct
1	satresp	737767.00	77.66
2	satadv	737725.00	77.66
3	satsoc	737722.00	77.66
4	satloc	737687.00	77.66
5	satchal	737649.00	77.65
6	satben	737636.00	77.65
7	satsal	737633.00	77.65
8	satind	737569.00	77.64
9	satsec	737405.00	77.63
10	jobins	642778.00	67.66

2.2 Data problems and solving

- Every person should only be surveyed once per year. there are 94,996 observations duplicated which have to be solved by distinct command.
- Variables are widely coded inconsistently. There are 52,277 missing values in Age variable, 52,297 values in Salary variable. salary has extreme values: median = 68k , but max = 9,999,999, likely censored)
- There are some typo errors in gender variable detailed in below table which caused inconsistency in the variable. by recoding the gender variable the problem has been solved.

	Female	Female	Male	Male
Initial number	260820	65415	460581	115458
Recoded error-free number	326235	0	576039	0

Table 2: Gender distribution Table

- The age variable consists of values between 23 and 99. Because the distance between the last age and the one before reveals a huge gap (24 years) and this category includes only 6 values, it could be considered as a potential missing values in between.

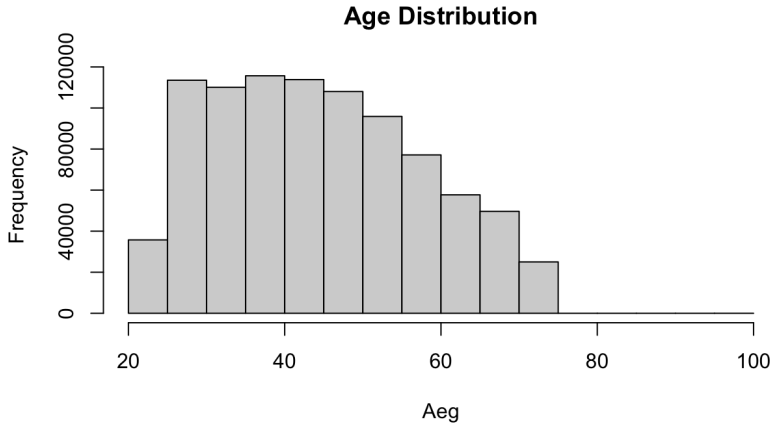

Figure 1: Age Distribution

- There are some typo errors in dgrdg(highest degree) variable detailed in below table which caused inconsistency in the variable. by recoding the gender variable the problem has been solved.

	Bachelor		Doctorate		Master		Professional	
	Initial	Recoded	Initial	Recoded	Initial	Recoded	Initial	Recoded
Count	280,112	350,788	237,134	296,088	174,098	217,423	30,341	37,811
Other	70,676	0	58,954	0	43,325	0	7,470	0

Table 3: Degree Distribution with Initial and Recoded Values

- There are also some typo errors in "ctzusin" variable which was recoded and solved. The type of typo errors categorized into 3 groups which can be seen in table below.

	No		Yes	
	Initial	Recoded	Initial	Recoded
Category 1	5,332	71811	62,384	830,463
Category 2	61,106	0	705,479	0
Category 3	5,373	0	62,600	0

Table 4: US citizen Distribution (Initial and Recoded)

- Under the "lfstat" (Employment status) there are several typo errors causing repeated errors(Employed, Unemployed and Not in the labor force) which was separated into two categories.
- The categorical data is recommended to be presented in Factor format. Therefore

	Employed		Not in the labor force		Unemployed	
	Initial	Recoded	Initial	Recoded	Initial	Recoded
Count	620,086	774,830	83,722	104,823	17,997	22,526
Other	154,744	0	21,101	0	4,529	0

Table 5: Employment Distribution with Initial and Recoded Values

3 Exercise III: Data Aggregation and Descriptive Statistics

In this section, we want to analyze job satisfaction patterns by educational attainment. Create a comprehensive table showing 2-3 different job satisfaction measures (e.g., overall job satisfaction, salary satisfaction, or advancement opportunities) by degree type and field of study.

3.1 Analyze job satisfaction patterns by educational attainment

3.1.1 creating job satisfaction dataset

In this part, we generate a specific dataset including person identifiers, Degrees and satisfaction measurement data for analyzing purpose.

3.1.2 cleaning dataset

In this part, we need to clean the data in terms of missing values, typo errors, duplicated values and any other misleading information. Because some values are in the same categories but with different lower/upper case characters, we have to standardize them and cut out the tailing and whitespaces first. Then we can filter to keep rows with AT LEAST ONE non-NA satisfaction measure.

3.1.3 Analyzing the satisfaction by educational status

In this section, we use the cleaned data to analyze the satisfaction variables(Job Satisfaction, Salary Satisfaction and Opportunity Satisfaction) by

the Educational Status. We can evaluate conditions sequentially and assign 4(very satisfied) to 1(very dissatisfied). Then calculate the percentage of Salary Satisfaction and Advancement Opportunity in that regards.

Degree	Job Satisfaction	Salary Satisfaction	Advancement Opp	Nunmber
Professional	3.51	30.55	30.46	22835
Doctorate	3.41	32.67	32.63	161828
Master's	3.36	38.02	37.92	124131
Bachelor's	3.30	33.53	33.57	186879

Table 6: Job Satisfaction Summary by Degree

A In the table above you can see

- **The average of Job Satisfaction Scores (1-4 Scale)**
  Professional degree holders report the highest job satisfaction (3.51), followed by Doctorate (3.41), Master's (3.36), and Bachelor's (3.30).
- The scores are relatively close, suggesting degree type has a modest (but not drastic) impact on overall satisfaction.
- **Salary Satisfaction percentage**
  Master's degree holders are the most satisfied with their salary (38.02), while Professional degree holders are the least satisfied (30.55).
- Surprisingly, higher degrees (Doctorate/Professional) do not correlate with higher salary satisfaction.
- **Advancement Opportunities percentage**
  Mirroring salary trends, Master's holders again report the highest satisfaction with advancement opportunities (37.92)
- The gap between degrees is small (30–38), suggesting advancement satisfaction is broadly similar across education levels.
- **Sample Size ("Opp" Column)**
  Bachelor's degree holders dominate the sample (186,879 responses), followed by Doctorate (161,828). The "Professional" category has the smallest sample (22,835), which may affect the reliability of its metrics.

- **Missing Data**
  The "Number" column is empty, which might indicate incomplete data or a placeholder for additional metrics (e.g., response rates).

3.2 Examining career-education alignment and its relationship to satisfaction.

3.2.1 Degree alignment with satisfaction

Stronger alignment between degree and job correlates with higher satisfaction across all measured factors.

- **Job Satisfaction (1-4 Scale)** Highest for closely related jobs (3.50), decreasing for somewhat related (3.30) and not related (3.10).
- **Salary & Advancement Satisfaction** Both metrics follow the same trend: Closely related: 34.8% satisfied with salary, 34.7% with advancement. Not related: 32.5-32.6% satisfied (2-3% lower).
- **Sample Size (N)** Majority of respondents work in closely related jobs (289,856 vs. 61,572 "not related").

Job related degree	Job Satisfact(1-4)	%Satisfy Salary	%Satisfy by Advancement	N
closely related	3.50	34.80	34.70	289856
somewhat related	3.30	33.90	33.90	119480
not related	3.10	32.50	32.60	61572

Table 7: Job Satisfaction Summary related to degree

3.2.2 Career-education alignment

- **Highest Alignment** Health-related fields dominate with 78.9% of graduates in closely related jobs, far surpassing other fields. Civil engineering follows at 70.1%, reflecting strong industry alignment.
- **STEM vs. Non-STEM** STEM fields (e.g., computer science, engineering) show moderate alignment (59-62%), with notable exceptions like electronics/communication engineering (59.8%). Non-science fields (e.g., psychology, "other non-science") still maintain ¿60% alignment, suggesting broader applicability.

- **Lowest "Not Related" Rates** Civil engineering (5.4%) and health fields (5.8%) have the fewest graduates in unrelated jobs. "Other non-science" fields have the highest mismatch (15.4%).
- **Somewhat Related" Trends** Electronics/communication engineering has the highest percentage (28.2%) in somewhat related roles, indicating potential skill transferability.

Field of Study	Closely Related	Not Related	Somewhat Related
health-related fields	78.90	5.80	10.40
civil engineering	70.10	5.40	19.20
other non-science and engineering	63.30	15.40	16.40
other science and engineering-related	61.90	11.70	21.60
psychology	61.90	13.80	19.20
computer and mathematical sciences	61.80	9.40	23.90
electronics and comm engineering	59.80	6.90	28.20
other physical sciences	58.70	13.70	22.50

Table 8: Top 8 Fields by Percentage Closely Related to Degree

3.3 Analyzing work activities and educational background

3.3.1 distribution of work activities by degree characteristics. Higher degrees (Doctorate/Master's) correlate with specialized roles (RD, Teaching), while Bachelor's lean toward broader applications (Management, Tech).

- **Research & Development (RD) Activity**
  Doctorate holders lead in RD involvement (60.9%), nearly double the rate of Bachelor's holders (34.2%). Professional degree holders show the lowest RD participation (14.6%).
- **Computer Application**
  Highest among Bachelor's (20.6%) and Master's (19.0%) graduates, likely reflecting tech-focused roles. Very low for Professional (2.0%) and Doctorate (9.5%), suggesting non-technical specializations.

- **Management Roles**
  Common across all degrees (35–53%), peaking for Bachelor's (53.3%). Implies management is a frequent career path regardless of education level.

- **Teaching**
  Doctorate holders are most likely to teach (30.3%), aligning with academic careers. Bachelor's have the lowest teaching involvement (9.9%).

- **Sample Size (n)**
  Bachelor's dominate the dataset (350,788 respondents), while Professional degrees are rarest (37,811).

Degree Type	RD(%)	Computer Application(%)	Management(%)	Teaching(%)	n
doctorate	60.9	9.5	35.8	30.3	296088
master's	39.0	19.0	47.9	17.3	217423
bachelor's	34.2	20.6	53.3	9.9	350788
professional	14.6	2.0	47.6	15.1	37811

Table 9: Education Attainment by Working Activity

3.3.2 correlations between education variables and work activity engagement. Degree fields strongly predict career sector specialization—tailor education and guidance accordingly.

- **Top RD Fields**
  Mechanical engineering, physics, and electrical engineering lead in research (62-63%).

- **Tech Roles**
  Electrical/electronics engineering has the highest computer application (31.5%).

- **Management Focus**
  Civil engineering dominates management (59%).

- **Teaching**
  Biological sciences and physics/astronomy have the most teaching roles (20-24%).

- **Sample Size**
  Biological sciences has the largest respondent pool (106K), reflecting strong STEM participation.

Field of highest degree	RD(%)	Computer(%)	Managmnt(%)	Teach(%)	n
mechanical engineering	62.6	15.4	46.6	5.5	37930
physics and astronomy	62.6	23.9	29.9	21.6	29058
electrical& com engineering	62.0	31.5	39.2	6.4	52723
chemical engineering	60.8	12.9	44.3	6.1	20654
chemistry, expt biochemistry	57.6	7.7	39.7	17.4	44001
other engineering	55.9	21.4	46.2	8.4	67774
other related sciences	54.7	14.9	39.4	19.6	23293
biological sciences	54.2	5.9	39.7	23.8	106104
civil engineering	49.1	14.6	59.0	6.4	28183
other biological, agri sciences	48.2	7.9	47.0	18.6	36078

Table 10: Top 10 Fields by Working Activity

4 Regression Analysis

4.1 Estimate a linear regression model predicting overall job satisfaction

- **Education Matters, But Not Always**
  Professional degrees boost satisfaction, but PhDs/master's add little. Professional degrees (e.g., MBA, JD): +0.130 (highest impact). (Doctorate: +0.053 , Master's: +0.018) Professional degree holders (e.g., lawyers, executives) report much higher satisfaction, likely due to higher pay/status.
  Doctorates and master's degrees show minimal gains over bachelor's degrees (reference group), suggesting diminishing returns for advanced education in job satisfaction.

- **Job-Education Relationship**
  Misalignment reduces satisfaction more than any other factor. Not related (Ref): -0.336 , Somewhat related: -0.271. Jobs unrelated to education have much lower satisfaction (by 0.3+ units). Strong alignment between education and work is critical for satisfaction.

- **Training Participation**
  Employers can boost satisfaction by investing in employee development. Received training (yes): +0.077. Training is linked to moderately higher satisfaction, likely due to skill development and employer investment.

- **Field of Study**
  ◇ **Positive Effects**: Health-related fields: +0.056. Economics: +0.060. These fields may align with stable, high-demand jobs (e.g., healthcare, finance).
  ◇ **Negative Effects**: Sociology/Anthropology: -0.040. Electrical Engineering: -0.035. Potential mismatch between education and job market demands (e.g., engineering roles may involve high stress).

- **Demographic Gaps Exist**
  ◇ **Age**: Coefficient: +0.007, Each additional year of age is associated with a very slight increase in job satisfaction (0.007 units). Older workers may feel more settled or experienced in their roles.
  ◇ **Gender**: Male (Ref: Female): +0.033, Men report slightly higher satisfaction than women, suggesting potential gender disparities in workplace experiences (e.g., pay, promotion opportunities, or work-life balance).
  ◇ **Race/Ethnicity**: White (Ref): +0.205,(highest satisfaction) Under-represented minorities: +0.138 Other: -0.116. White workers report the highest satisfaction, while "other" racial groups (unspecified) report lower satisfaction. Under-represented minorities fall in between. This may reflect systemic inequities or cultural differences in workplace treatment.

- **Other Considerations**
  ◇ Educational Attainment Paradox: Higher degrees do not guarantee higher satisfaction, possibly due to mismatched expectations or limited high-status job availability.

Table 11: linear regression model predicting overall job satisfaction

Function	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	2.4374	0.0103	237.53	0.0000
age	0.0075	0.0002	46.99	0.0000
gender	0.0333	0.0040	8.25	0.0000
racethother	-0.1164	0.1996	-0.58	0.5598
racethunder-represented minorities	0.1380	0.0064	21.65	0.0000
racethwhite	0.2054	0.0053	39.10	0.0000
dgrdgdoctorate	0.0532	0.0049	10.83	0.0000
dgrdgmaster's	0.0176	0.0049	3.58	0.0003
dgrdgprofessional	0.1297	0.0102	12.68	0.0000
ndgmedchemical engineering	0.0020	0.0135	0.14	0.8852
ndgmedchemistry, except biochemistry	-0.0279	0.0101	-2.75	0.0059
civil engineering	0.0222	0.0119	1.86	0.0626
computer and mathematical sciences	-0.0263	0.0082	-3.19	0.0014
economics	0.0603	0.0119	5.04	0.0000
electrical, elect and commu engineering	-0.0349	0.0095	-3.66	0.0003
health-related fields	0.0560	0.0089	6.28	0.0000
management and administration	-0.0480	0.0107	-4.50	0.0000
mechanical engineering	-0.0334	0.0107	-3.13	0.0017
missing	0.0321	0.7600	0.04	0.9664
other biological, agricultural, environmental	0.0106	0.0114	0.93	0.3504
other engineering	0.0099	0.0089	1.11	0.2669
other non-science and engineering	0.0015	0.0092	0.17	0.8682
other physical and related sciences	0.0137	0.0129	1.06	0.2889
other science and engineering-related	-0.0430	0.0130	-3.32	0.0009
other social sciences	0.0125	0.0121	1.04	0.3004
physics and astronomy	0.0197	0.0119	1.66	0.0978
political and related sciences	0.0286	0.0127	2.25	0.0246
psychology	0.0116	0.0083	1.39	0.1656
sociology and anthropology	-0.0404	0.0111	-3.63	0.0003
ocedrlpnot related	-0.3356	0.0057	-58.56	0.0000
ocedrlpsomewhat related	-0.2712	0.0043	-62.50	0.0000
wktrni	0.0765	0.0038	20.24	0.0000

4.2 Model predicting salary satisfaction by degree

Regression analysis of salary satisfaction. Salary satisfaction is strongly tied to field of study, benefits, and demographic factors, with STEM careers and professional degrees offering the highest returns. Higher satisfaction is related to STEM fields, professional degrees, and strong benefits. Age, gender, and race also matter.

- **Demographics Matter**
  Older employees and men report higher satisfaction, while white employees are more satisfied than underrepresented minorities.
- **Education Pays Off**
  Professional degree holders show the highest satisfaction, while doctorates report slightly lower satisfaction than bachelor's degrees (reference group).
- **Field-Specific Trends**
  STEM fields (e.g., chemical engineering, computer science) correlate with higher satisfaction, while non-STEM fields (e.g., social sciences) often show lower satisfaction.
- **Benefits Impact**
  Satisfaction spikes for those "very satisfied" with benefits (satben), highlighting the importance of compensation packages.
- **Statistical Significance**
  Most estimates are highly significant (p ¡ 0.05), except for master's degrees and some fields of social science.

Table 12: Salary Satisfaction Summary related to degree

Function	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5402	0.0155	99.56	0.0000
age	0.0061	0.0002	27.19	0.0000
gender	0.0484	0.0059	8.23	0.0000
racethunder-represented minorities	0.0428	0.0087	4.90	0.0000
racethwhite	0.1000	0.0074	13.49	0.0000
dgrdgdoctorate	-0.0286	0.0070	-4.07	0.0000
dgrdgmaster's	0.0097	0.0070	1.38	0.1686
dgrdgprofessional	0.1309	0.0153	8.54	0.0000
ndgmedchemical engineering	0.1866	0.0201	9.29	0.0000
ndgmedchemistry, except biochemistry	0.0662	0.0153	4.32	0.0000
ndgmedcivil engineering	0.0523	0.0179	2.93	0.0034
computer and math sciences	0.1055	0.0122	8.67	0.0000
ndgmedeconomics	0.1047	0.0170	6.16	0.0000
electrical, elect & commu engineering	0.1523	0.0146	10.41	0.0000
ndgmedhealth-related fields	0.0773	0.0123	6.30	0.0000
management and administration	0.1183	0.0157	7.55	0.0000
ndgmedmechanical engineering	0.1383	0.0157	8.81	0.0000
other bio, agri, envirlife sciences	0.0129	0.0176	0.73	0.4637
ndgmedother engineering	0.1171	0.0130	9.01	0.0000
dother non-science and engineering	-0.0284	0.0135	-2.10	0.0355
dother physical & related sciences	0.1028	0.0196	5.25	0.0000
dother science and engineering-related	-0.0200	0.0175	-1.14	0.2525
ndgmedother social sciences	-0.0681	0.0192	-3.54	0.0004
ndgmedphysics and astronomy	0.0660	0.0185	3.58	0.0003
political and related sciences	-0.0033	0.0171	-0.19	0.8471
ndgmedpsychology	-0.0134	0.0122	-1.10	0.2708
dsociology and anthropology	-0.0520	0.0165	-3.15	0.0016
satbensomewhat satisfied	0.2295	0.0087	26.39	0.0000
satbenvery dissatisfied	0.3810	0.0123	31.05	0.0000
satbenvery satisfied	1.0953	0.0088	124.57	0.0000

4.3 Career advancement satisfaction Summary related to degree and work activity

The study of 178K professionals reveals that job-education alignment is the strongest predictor - satisfaction plummets by 0.34 points when work is unrelated to one's degree, while professional degree holders gain 0.19 points, underscoring the career value of specialized education.

- **Education Effects**
  The doctorate advantage (+0.09) contrasts with master's holders' slight disadvantage (-0.02), revealing nuanced returns on graduate education that likely reflect differing career expectations and opportunities.

- **Notable Effects**
  Professional degrees (+0.19) and training participation (+0.13) boost satisfaction, while jobs unrelated to education sharply reduce it (negative 0.34).

- **Demographics**
  Women report lower satisfaction (negative 0.06) than men, and white employees score higher (+0.05) than underrepresented minorities.

- **Workplace Dynamics**
  Company size shows a clear inverse relationship - satisfaction drops 0.23-0.29 points in larger firms (500+ employees), suggesting smaller organizations may offer better advancement experiences, while training participation boosts satisfaction by 0.13 points.

- **Age Impact**
  Older employees show slightly lower satisfaction (negative 0.002 per year), though the effect is small in magnitude.

Table 13: Career advancement satisfaction Summary related to degree

Function	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1284	0.0173	180.90	0.0000
age	-0.0018	0.0002	-7.96	0.0000
Female	-0.0574	0.0055	-10.46	0.0000
racethunder-represented minorities	-0.0126	0.0085	-1.48	0.1379
racethwhite	0.0481	0.0072	6.68	0.0000
dgrdgdoctorate	0.0861	0.0070	12.24	0.0000
dgrdgmaster's	-0.0186	0.0067	-2.80	0.0052
dgrdgprofessional	0.1888	0.0142	13.33	0.0000
ocedrlpnot related	-0.3387	0.0086	-39.54	0.0000
ocedrlpsomewhat related	-0.1887	0.0063	-29.88	0.0000
wktrni	0.1256	0.0055	23.00	0.0000
waprsmmanagement and administration	0.0662	0.0113	5.85	0.0000
waprsmother	-0.0719	0.0118	-6.09	0.0000
waprsmresearch and development	-0.0128	0.0114	-1.13	0.2596
waprsmteaching	-0.0890	0.0128	-6.96	0.0000
emsize100-499 employees	-0.2878	0.0110	-26.18	0.0000
emsize1000-4999 employees	-0.2637	0.0107	-24.68	0.0000
emsize11-24 employees	-0.1802	0.0152	-11.85	0.0000
emsize25-99 employees	-0.2781	0.0122	-22.75	0.0000
emsize25000+ employees	-0.2296	0.0095	-24.15	0.0000
emsize500-999 employees	-0.2880	0.0131	-22.06	0.0000
emsize5000-24999 employees	-0.2349	0.0106	-22.17	0.0000

5 Data Visualization

5.1 the relationship between salary and overall job satisfaction

- **Salary-Satisfaction Relationship** The positive slopes of regression lines suggest higher salaries generally correlate with higher job satisfaction across all degree levels.The strength of this relationship varies by degree type (steeper slopes for Doctorate/Professional degrees).

- **Degree-Level Differences**
- Professional degree holders (red line) likely show highest satisfaction at comparable salaries strongest salary-satisfaction relationship (steepest slope). Bachelor's degrees (blue) may show Flatter slope, suggesting salary increases matter less for their satisfaction lower satisfaction at higher salaries compared to advanced degrees.
- **Potential Insights** The "Professional" group's curve suggests their satisfaction is more sensitive to salary changes than other groups. Master's and Doctorate lines appear closer, possibly indicating similar satisfaction drivers. At lower salaries (¡$50K), satisfaction differences between degrees are minimal, but they diverge sharply at higher salaries.

Figure 2: Relationship between salary and overall job satisfaction