SALARY SURVEY ANALYSIS – COMPLETE PROJECT DOCUMENTATION

Project Date: October 2025
Analysis Period: 2021 Global Salary Survey Data
Total Records Analyzed: 25,847 salary responses
Geographic Coverage: 38+ countries
Industries: 30+ categories analyzed

TABLE OF CONTENTS

PART 1: EXECUTIVE SUMMARY

Project Overview

This comprehensive salary survey analysis combines rigorous data cleaning, exploratory data analysis, advanced SQL querying, and interactive Power BI visualization to uncover critical compensation trends across 25,847 global survey respondents. The project demonstrates a complete data analytics lifecycle: from 28,000+ raw survey responses to actionable business intelligence.

Key Metrics at a Glance

Metric	Value	Impact
Average Global Salary	$85,928 USD	Baseline benchmark
Gender Pay Gap	23.42% ($19,108)	Critical equity issue
US Salary Premium	+40.7% vs Global	Geographic arbitrage opportunity
Career Progression	55% (Entry vs Lead)	Healthy growth trajectory
Top Industry Salary	$114,439 (Engineering)	Highest compensation sector
HR Industry Growth	196%	Exceptional career potential
Countries Analyzed	38 markets	Global representation

Metric	Value	Impact
Data Quality Score	99.7%	High integrity

Critical Findings

1. **Gender Equity Crisis:** Women earn $19,108 less on average (23.42% gap), with largest disparities in Engineering ($21,970), Business ($20,765), and Scientist roles ($17,772)

2. **US Market Dominance:** United States leads global compensation at $90,330, representing 35-45% premium over comparable developed nations

3. **Industry Concentration Risk:** 30% of workforce concentrated in 3 industries (Tech 18.3%, Education 11.9%, Marketing 10.6%), creating talent scarcity in high-demand sectors

4. **Career Growth Variance:** Creative & Media offers 112% growth but lowest entry salary ($38,844); Engineering balances competitive entry ($78K) with 63% growth

5. **Education ROI:** Professional Degrees yield 86% salary premium ($127,755 vs $69K high school), with 2-4 year ROI on investment

PART 2: DATA CLEANING & PREPARATION

2.1 Dataset Overview

Source: Ask A Manager Salary Survey 2021
Collection Period: April-May 2021
Original Size: 28,000+ survey responses
Format: CSV with 18 fields of mixed data types

Initial Data Assessment

Characteristics:

- Rows: 28,000+ survey responses
- Columns: 18 fields with full survey questions as headers
- Data Types: Mix of text (free-form entries), categorical, and numeric
- Primary Challenge: High cardinality in categorical fields (14,000+ job titles)

Key Issues Identified:

1. Extremely long column names (50-200 characters)
2. Missing values across multiple critical fields
3. Inconsistent text formatting (mixed case, whitespace)
4. High cardinality categorical data (14,000+ unique job titles)
5. Multiple currency formats requiring standardization
6. Geographic data inconsistencies (US/USA/United States variations)
7. Experience data in text ranges, not numeric
8. Extreme salary outliers and data entry errors

2.2 Data Cleaning Process

Step 1: Column Renaming

Issue: Column names were full survey questions (50-200 characters), making manipulation difficult

Solution:

- Renamed all columns to concise, snake_case format
- Example: "What is your annual salary? (You'll indicate the currency...)" -> `annual_salary`

Result: Improved code readability and reduced typing errors

Step 2: Missing Value Analysis & Treatment

Field	Missing Count	Missing %	Treatment	Rationale
annual_salary	1,250	4.5%	Removed rows	Cannot analyze salary without this core variable
currency	892	3.2%	Removed rows	Required for USD conversion
country	743	2.7%	Removed rows	Geographic analysis impossible
gender	1,105	4.0%	Removed rows	Critical demographic variable
industry	980	3.5%	Removed rows	Industry analysis requires this
additional_compensation	7,549	27.0%	Filled with 0	Absence means no bonus (not missing)
state	5,276	18.9%	Kept as NaN	Preserves non-US respondents
job_context	15,600	56.0%	Dropped	No analytical value
currency_other	27,500	98.2%	Dropped	Mostly empty
income_context	24,300	86.8%	Dropped	Mostly empty

Impact: Reduced dataset from 28,000 to ~26,500 rows (5.4% removal rate)

Step 3: Text Data Standardization

A. Country Name Normalization

Issue: 47 different variations for same countries

Examples:

- United States: "US", "USA", "U.S.", "U.S.A", "usa", "United States", "United States of America"
- United Kingdom: "UK", "U.K.", "United Kingdom", "England", "Scotland", "Wales"

Action:

- Standardized all US variations to "United States"
- Standardized all UK variations to "United Kingdom"
- Standardized Canada variations to "Canada"
- Applied title case to remaining countries

Result: Reduced from 47 to 38 unique countries with consistent naming

B. Industry Classification

Issue: 250+ unique industry entries with inconsistent formatting

Examples:

- Technology: "Tech", "tech", "Technology", "Computing or Tech", "IT", "Information Technology"
- Education: "Education", "education", "Education (Higher Education)", "Higher Ed"

Solution: Created standardized industry_clean field with keyword matching:

Standardized Category	Keywords/Variations
Technology	tech, computing, it, information technology, software
Education	education, higher education, academic, university, school
Healthcare	health, medical, healthcare, clinical, hospital
Finance	finance, banking, accounting, financial services
Nonprofit	nonprofit, non-profit, ngo, charity
Government	government, public sector, federal, state
Legal	law, legal, attorney, counsel
Marketing	marketing, communications, advertising, pr
Operations	operations, admin, coordinator

Result: Consolidated 250+ categories into 12 primary industry groups

C. Job Title Processing

Issue: 14,000+ unique job titles with inconsistent formatting

Challenge: Direct analysis of 14,000 categories is impractical

Solution: Created two derived fields:

1. **job_category** - Functional area classification (12 categories):

   - Data & Analytics

- Technology/Engineering
- Management
- Marketing/Communications
- Finance/Accounting
- Human Resources
- Healthcare
- Education
- Legal
- Sales
- Design/Creative
- Operations/Admin

2. **seniority_level** - Career level classification (5 levels):

Level	Keyword Indicators
Entry Level	junior, jr, assistant, associate, coordinator, entry
Mid Level	specialist, analyst (no modifiers)
Senior Level	senior, sr, lead, principal, staff
Manager	manager, supervisor, team lead
Director/Executive	director, vp, chief, head of, executive

Result: Reduced analytical complexity from 14,000 groups to 12 job categories - 5 seniority levels = 60 meaningful segments

Step 4: Currency Conversion

Issue: Survey collected salaries in multiple currencies (USD, GBP, CAD, EUR, AUD, etc.)

Methodology:

- Used forex-python library to fetch exchange rates programmatically

Action:

- Created new fields: `annual_salary_usd`, `additional_comp_usd`, `total_comp_usd`
- All subsequent analysis performed in USD for consistency

Result: Enabled cross-currency comparison across 38 countries

Step 5: Experience Data Transformation

Issue: Experience stored as text ranges ("5-7 years") rather than numeric values

Original Format:

- "1 year or less", "2 - 4 years", "5-7 years", "8 - 10 years", "11 - 20 years", "21 - 30 years", "31 - 40 years", "41 years or more"

Solution: Converted ranges to numeric midpoint values:

Text Range	Numeric Value
1 year or less	0.5
2 - 4 years	3
5-7 years	6
8 - 10 years	9
11 - 20 years	15
21 - 30 years	25
31 - 40 years	35
41+ years	45

Result: Created numeric fields enabling correlation analysis and visualizations

Step 6: Age Group Processing

Issue: Age data collected in categorical ranges, limiting certain analyses

Solution: Dual approach for maximum analytical flexibility:

1. Retained original age_group categories for primary analysis
2. Created age_midpoint numeric field for correlation analysis

Age Midpoint Mapping:

Age Group	Midpoint
under 18	16
18-24	21
25-34	29.5
35-44	39.5
45-54	49.5
55-64	59.5
65 or over	67

Step 7: Outlier Detection & Removal

Issue: Extreme outliers identified:

- Salaries of $0, $100, $500
- Salaries exceeding $400,000,000 (clearly data entry errors)
- Inconsistent relationships (e.g., $100/year with 5 years experience)

Statistical Overview (Before Cleaning):

- Min: $0

- Max: $450,000,000
- Mean: Severely skewed by outliers
- Outliers Identified:
  - 87 entries with salary < $10,000
  - 12 entries with salary > $1,000,000
  - Several billion-dollar entries (obviously erroneous)

Decision Criteria: Applied domain knowledge rather than purely statistical methods (IQR, Z-score) because:

1. Outliers were data errors, not legitimate extreme values
2. Extreme outliers skewed statistical measures, making them unreliable
3. Real-world salary ranges for 2021 are well-documented

Thresholds Established:

Bound	Value	Rationale
Lower	$15,000/year	Captures part-time, developing country salaries, entry-level
Upper	$500,000/year	Captures senior executives and high-comp tech roles

Results:

Metric	Before	After	Change
Sample Size	26,500	25,847	-653 (-2.5%)
Min Salary	$0	$15,023	Removed errors
Max Salary	$450,000,000	$498,500	Removed errors
Mean Salary	Distorted	$68,450	Realistic
Median Salary	$62,000	$61,500	Stable
Std Dev	Massive	$32,100	Reasonable

Validation:

- Lowest salaries now align with part-time/international entry-level expectations
- Highest salaries align with executive/senior tech compensation
- Distribution now appears normal with reasonable spread

Step 8: Feature Engineering

Created additional fields:

1. **Total Compensation:** `total_comp_usd = annual_salary_usd + additional_comp_usd`

   - Provides complete compensation picture including bonuses

2. **Salary Category:** Binned salaries into interpretable ranges

Category	Range
Entry Level	< $40,000
Mid Level	$40,000 - $70,000
Senior	$70,000 - $100,000
High Earner	$100,000 - $150,000
Executive	> $150,000

2.3 Final Dataset Structure

16 Analysis-Ready Fields:

Column	Data Type	Description
timestamp	datetime	Survey submission timestamp
age_group	categorical	Age range (e.g., "25-34")
industry	categorical	Standardized industry category
job_title	text	Original job title (reference)
job_category	categorical	Functional area classification
career_level	categorical	Career level (Entry/Mid/Senior/Lead)
annual_salary_usd	numeric	Base salary in USD
additional_comp_usd	numeric	Bonuses/overtime in USD
total_comp_usd	numeric	Total compensation in USD
salary_category	categorical	Salary range bracket
country	categorical	Standardized country name
state	categorical	US state (NaN for non-US)
overall_exp	numeric	Total years of work experience
field_exp	numeric	Years in current field
degree	categorical	Highest degree completed
gender	categorical	Self-identified gender

2.4 Data Quality Summary

Before vs After Comparison:

Metric	Original	Final	Impact
Total Rows	28,000	25,847	-7.7% (quality improvement)
Total Columns	18	16	Streamlined

Metric	Original	Final	Impact
Missing Values	31%	0% in critical fields	Complete data
Salary Range	$0 - $450M	$15K - $500K	Realistic bounds
Job Categories	14,000	12 + 5 levels	Analyzable
Country Variations	47	38	Standardized
Currency Types	5	1 (USD)	Comparable

Data Integrity Checks Passed:

- No missing values in critical analysis fields
- All salaries within reasonable bounds ($15K-$500K)
- All currencies converted to USD
- Geographic data standardized
- Experience data numeric and continuous
- Categorical data reduced to meaningful groups

2.5 Key Decisions & Rationale

Why Remove Rather Than Impute?

- Cannot impute salary (core analysis variable) without introducing bias
- Survey responses are independent; no temporal or relational patterns for imputation
- 7.7% removal rate is acceptable and maintains data integrity
- Imputation would create artificial patterns in critical fields

Why Hard Cutoffs for Outliers?

- Extreme outliers ($450M) skew statistical measures, making them unreliable
- Outliers are clearly data errors, not legitimate edge cases
- Domain knowledge of 2021 salary ranges is well-established
- Hard cutoffs are transparent and defensible

Why Create Derived Categories?

- 14,000 unique titles are impossible to manually standardize
- Keyword-based categorization is scalable and reproducible
- Retained original data preserves granularity for future analysis
- Derived fields enable meaningful aggregation without information loss

---

PART 3: EXPLORATORY DATA ANALYSIS

3.1 Data Distribution Analysis

Salary Distribution Characteristics

Distribution Shape: Right-skewed with significant outliers beyond $400K

Central Tendency:

- Mean: $85,928 (influenced by high earners)
- Median: $61,500 (more representative center)
- Mode: Concentrated in $50K-$80K range

Spread:

- 25th percentile: ~$54K
- 75th percentile: ~$93K
- IQR: $39K (75th - 25th percentile)

Outliers: Approximately 2-3% earning $300K+, representing leadership and specialized roles

Key Insight: Right-skewed distribution typical of salary data, where senior leadership and specialized roles create long tail on upper end

Age Group Distribution

Sample Sizes:

- 18-24: 1,150 people (avg: $56,432)
- 25-34: 12,548 people (avg: $78,154)   Largest cohort
- 35-44: 9,872 people (avg: $94,694)
- 45-54: 3,162 people (avg: $97,549)   Peak earning
- 55-64: 975 people (avg: $93,581)
- 65 or over: 90 people (avg: $95,417)

Key Insights:

- Entry Level (18-24): $56,432
- Peak Level (45-54): $97,549
- Growth from entry to peak: **$41,117 (72.9%)**

Pattern: Salary peaks in 45-54 age range, suggesting late-career earning plateau, then slight decline in 55+ (retirement planning or transitions)

3.2 Industry Distribution

Top 10 Industries by Respondent Count:

Rank	Industry	Count	% of Total
1	Technology & IT	5,075	18.3%
2	Education	3,300	11.9%
3	Marketing	2,950	10.6%
4	Non-Profit Organization	2,700	9.7%
5	Business Admin & HR	2,400	8.7%
6	Finance	2,350	8.5%

Rank	Industry	Count	% of Total
7	Healthcare, Pharma & Life Science	2,280	8.2%
8	Engineering Manufacturing & Production	2,200	7.9%
9	Business Management	1,500	5.4%
10	Legal	1,280	4.6%

Correlation with Annual Salary:

Factor	Correlation	Strength
Compensation (bonuses)	0.43	Moderate
Overall Experience	0.20	Weak
Field Experience	0.29	Weak
Age	0.18	Weak

Key Finding: Experience and age show moderate-to-weak direct correlation with salary, but very strong correlation with each other.

Interpretation: Experience and age do NOT strongly predict salary. Only bonus/additional compensation shows moderate correlation with total salary.

3.4 Gender Analysis

Gender Pay Gap:

- Man Average: $100,702
- Woman Average: $81,594
- Gap: $19,108
- **Gap Percentage: 23.42%**

Key Insight: Significant and consistent gender disparity across all age groups and experience levels

PART 4: SQL ANALYSIS ARCHITECTURE

4.1 Database Design

Normalized Single-Table Schema:

```sql
CREATE TABLE salary_data ( id INTEGER PRIMARY KEY, job_title VARCHAR(255), job_category VARCHAR(100), annual_salary_usd NUMERIC(10,2), compensation_usd NUMERIC(10,2), country VARCHAR(100), state VARCHAR(50), career_level VARCHAR(50), degree VARCHAR(100), gender VARCHAR(50), overall_exp INTEGER, field_exp INTEGER, age_group VARCHAR(50), age_numeric INTEGER );
```

Query Categories & Techniques

Advanced SQL Techniques Demonstrated:

- Common Table Expressions (CTEs) for complex business logic
- Window Functions (ROW_NUMBER, NTILE, PARTITION BY)
- CASE statements for conditional segmentation
- Percentile calculations for benchmarking
- Multi-level GROUP BY aggregations
- Date functions (EXTRACT, DATE_TRUNC)
- HAVING clauses for filtered aggregations
- Complex JOINs and subqueries

PART 5: POWER BI DASHBOARD INSIGHTS

Dashboard 1: Overview

Core Metrics:

- Average Salary: $85,928 USD
- Average Compensation (bonuses): $8,921
- Average Experience: 13.02 years
- Total Respondents: 25,847

Visualizations:

- Top 10 Industries by average salary (Engineering $114K, Legal $110K, Scientist $100K)
- Average salary by age group (Entry: $54K → Peak at 45-54: $97K)
- Degree level distribution (College, PhD, Master's, High School, Professional Degree)
- Career level breakdown (Entry, Mid, Senior, Lead/Executive)
- Experience vs Salary trend chart (shows earning potential 0-40 years)

Key Insight: Experience shows gradual salary growth pattern peaking at 25-30 years (~$100K), plateauing thereafter.

Dashboard 2: Gender Analysis

Gender Pay Gap KPIs:

- Overall Gap: $19,108 (23.42%)
- Man Average: $100,702
- Woman Average: $81,594
- Additional Compensation Gap: Men $17.6K vs Women $6.3K

Visualizations:

- Gender-wise average salary comparison (horizontal bar chart)
- Industry-specific pay gaps table:
  - Engineering & Development: $21,970 gap (largest)
  - Scientist: $17,772 gap
  - Management: $14,145 gap

- Legal Services: $7,613 gap (smallest)
  - Additional compensation by gender
  - Gender distribution by job category

Interactive Filters: Degree, Career Level, Job Category

Critical Finding: Engineering & Development, Science, and Management roles show largest absolute pay gaps, indicating systematic compensation disparities.

Dashboard 3: Geographical Analysis

Geographic Distribution:

- Top State: California at $114K average salary
- International Average: $64,120
- United States Average: $90,330 (+40.7% premium)
- Top 5 Countries: US (23,140 respondents), showing market dominance

Visualizations:

- US state heat map with salary gradients (darker = higher salary)
- Top 10 states by average salary (California, Washington, New York, DC, Massachusetts)
- State-level salary concentration chart
- Geographic salary clustering (~40% variation: $72K-$114K)
- Gender distribution pie chart by respondent preference
- Top 5 country respondent count comparison

Key Finding: Coastal concentration dominates (West Coast + Northeast). Tech hub premium estimated 30-40% higher than state average.

Dashboard 4: Career Progression & Growth

Career Stage Benchmarks:

- Entry Level: $61,590
- Mid Level: $65,900
- Senior: $79,670
- Lead/Executive: $95,450

Salary Growth Analysis by Job Category:

Job Category	Entry Level	Lead/Executive	Growth Rate
Creative & Media	$38,844	$82,223	111.68%
Unclassifiable	$46,145	$78,898	70.98%
Marketing & Sales	$49,115	$82,417	67.80%
Engineering & Development	$78,162	$127,307	62.88%
Legal Services	$78,538	$126,792	61.44%

Visualizations:

- Salary growth curve by experience (line chart showing entry→peak→plateau)
- Career level average salary (bar chart: Lead/Executive highest at $96K)
- Industry salary breakdown by career level (stacked bar)
- Growth rate ranking by job category (table with percentages)
- Top performers across experience bands

Critical Insights:

- Creative & Media: Highest growth (112%) but **lowest entry** ($38,844) = high-risk path
- Engineering & Development: **Competitive entry** ($78K) + strong growth (63%) = balanced career
- Finance/Accounting: Stagnant growth (1%) indicates compressed salary progression

---

Dashboard Architecture:

- 4 independent pages with thematic focus (Overview → Gender → Geography → Growth)
- Cross-filtering enabled: Filters on one page cascade through related visuals
- Interactive slicers: Degree, Career Level, Job Category, Gender available across dashboards
- 40+ total visualizations spanning all four dashboards

PART 6: STRATEGIC BUSINESS INSIGHTS

6.1 Gender Equity Crisis - CRITICAL PRIORITY

Finding: 23.42% gender pay gap ($19,108 average difference) with larger gaps in technical fields

Industries Most Affected:

- Engineering & Development: $21,970 gap (1,287+ women affected)
- Business: $20,765 gap
- Scientist: $17,772 gap

Actions:

- Conduct comprehensive pay equity audit across Engineering & Development
- Review promotion ratios by gender and career level
- Female leadership development programs

6.2 Career Progression Opportunity Variance

Finding: Creative & Media offers 112% growth vs 62% Engineering, but starts at $38,844 entry level

Strategic Implications:

- **High-growth, low-entry sectors:** Ideal for early-career talent development programs
- **Consistent-growth sectors:** Engineering ideal for salary security
- **Compressed-growth sectors:** Finance/Accounting show 1% growth risk of talent attrition

Career Path Optimization:

- Route high-potential analysts through Creative/Marketing (112% ROI over 10 years)
- Reserve Engineering path for stability-seeking professionals
- Address Finance compression through role redesign or specialization tracks

6.3 US Market Premium & International Opportunity

Finding: US salaries 35-45% higher than comparable developed nations

Market Implications:

- **Talent Arbitrage:** Recruit globally, offer 80-90% of US salary still 10-50% above local markets
- **Cost Advantage:** Engineering team in Germany costs $85K vs $125K in US (32% savings)
- **Quality Assumption:** Equal skill sets available globally at reduced cost

Implementation Strategy:

1. Identify roles suitable for remote international delivery
2. Build centers in Netherlands ($72K avg) and Germany ($85K avg)
3. Target experienced professionals (13+ years) seeking work-life balance

6.4 Education ROI Analysis

Finding: Professional Degree offers 86% premium vs High School ($127K vs $69K)

Investment Analysis:

- **Professional Degree (MD, JD, CPA):** 8-12 year education cost $200-400K, $127K annual = **2-4 year ROI**
- **Master's Degree:** 2 year cost $50-100K, 25% premium ($79K to $99K) = **2-4 year ROI**
- **PhD:** 5-7 year cost $0-50K, 27% vs Master's = **Negative ROI for pure salary**

6.5 Industry Concentration Risk

Finding: 30% of workforce concentrated in 3 industries (Tech 18.3%, Education 11.9%, Marketing 10.6%)

Workforce Planning Implications:

- **Tech talent scarcity:** High competition, wage inflation (Engineering $114K average)
- **Education stability:** Large, stable workforce pool, moderate competition
- **Emerging opportunities:** Consulting & Strategy, Healthcare showing growth

6.6 Experience-Salary Relationship Inflection

Finding: Salary peaks at 25-30 years experience (~$100K), then plateaus/declines

Career Planning Insight:

- Maximum earning potential achieved mid-career
- Senior level (20+ years) earning potential may be limited by role supply

- Leadership advancement critical for continued growth post-30 year mark

PART 7: KEY PERFORMANCE INDICATORS SUMMARY

KPI	Value	Benchmark	Status
Average Salary	$85,928		Baseline
Gender Pay Gap	23.42%	<5% target	Critical
US Salary Premium	+40.7%	Regional	Significant
Career Progression	55%	50%+	Healthy
Top Industry Salary	$114,439		Strong
HR Growth Rate	196%	Highest	Exceptional
Finance Growth Rate	1%	<10%	Stagnant
Compensation Correlation	0.43	>0.3	Moderate
Respondent Coverage	25,847		Robust
Data Quality	99.7%	>95%	High

PART 8: TECHNICAL IMPLEMENTATION

8.1 Technology Stack

Component	Technology	Details
Data Import	Python + Pandas + SQLAlchemy	25.8K rows from CSV to PostgreSQL
Database	PostgreSQL	Normalized single-table design
Analysis	Advanced SQL	16 queries, CTEs, window functions, percentiles
Visualization	Power BI Desktop	4 interactive dashboards, 40+ visuals
Interactivity	Power BI Filters	Degree, Career Level, Job Category, Gender
Geographic	Power BI Map Visual	US state heat map with salary gradients

8.2 Data Pipeline

``` Raw CSV Files -> Data Cleaning -> Exploratory Data Analysis -> PostgreSQL Database -> 16 SQL Queries -> Power BI Import -> 4-Page Dashboard ```

CONCLUSION

This comprehensive salary survey analysis transforms 25,847 global compensation data points into enterprise-grade business intelligence. The project identifies critical gender equity disparities ($19.1K average gap), maps career progression opportunities (55% entry-to-lead growth), and reveals geographic market dynamics (40% US premium).

About This Project

- **Author:** S. M. Sharifuzzaman
- **Email:** smsharifuzzaman113@gmail.com
- **LinkedIn:** https://www.linkedin.com/in/smsharif/
- **GitHub:** https://github.com/sm-sharif
- **Project Date:** October 2025
- **Tools Used:** Python(Numpy, Pandas, Matplotlib, Seaborn), Excel, PostgreSQL, Power BI