DELL Technologies

Website: https://www.dell.com/en-us

BUSINESS REPORT

Antonio Dulalia / adulali2@calstatela.edu
Jammal Losiea / jloisea@calstatela.edu
Ricardo Martinez / rmarti115@calstatela.edu

DELL Technologies

TABLE OF CONTENTS

Executive Summary

In 1984, Michael Dell began Dell Inc. in a Texas University dorm room now formally known as Dell Technologies. Dell Inc. created a massive disruption in the technology sector with its direct sales model of "build to order" and "configure to order." Dell Inc. also took pride in its innovative service solutions and technological supply chaining. Dell Inc. began to be known for its industry-leading and environmentally impactful innovations.

In the late 2000s, Dell Inc. started a business strategy revolving around its packaging. Dell Inc. called this strategy the "3 Cs Strategy." This method focuses on their package's size and shape (Cube), material and choice (Content), and recyclability (Curb). Dell Inc. began to implement this strategy by producing packaging materials made of plastics found littered the oceans of Caribbean islands. They did not stop there. Dell Inc. replaced petroleum-derived foam with bamboo harvested near its manufacturing facilities. They also began to grow packaging cushions made from mushrooms which allowed them to reduce the size of their packaging footprint.

To double down on this mission and commitment to the environment, Dell Technologies created a plan called the "2020 Legacy of Good." This plan outlines the company's promise to the environment by adding ten times the amount of "good" compared to technology's current footprint. This initiative expanded to Dell Technologies' supply chain, ensuring its partner companies' consistent, transparent environmental and social stewardship. "2020 Legacy of Good" eventually led Dell Technologies towards the Ocean Plastics Initiative.

Dell Technologies' approach to the Ocean Plastics Initiative includes four phases: (1) initial assessment, (2) Haiti pilot supply chain, (3) Asia-based supply chain, and (4) scaling. The initial assessment allowed Dell Technologies to recognize the exceedingly expensive, inefficient, and unreliable strategy method of collecting plastic from the ocean, which reassessed their strategy by focusing on preventing plastic from reaching the sea in the first place. The Haiti pilot supply chain phase of this initiative was a

success. It proved the technical viability of an ocean plastics supply chain, albeit with many improvements needed before it becomes a reality. This initiative's Asia-based supply chain phase was implementing the ocean plastics supply chain in South Asia. The two primary reasons for this are that it shortens lead time and logistics costs as it moves through Dell's packaging and product manufacturing sites in China. Second, it is one of South Asia's highest mismanaged waste areas.

The Ocean Plastics Initiative is one of Dell's ambitious international projects. Dell's ambitious goal inspired our team to present Dell with our project. The project we have come up with is to increase the diversity in Dell's workforce, specifically in Texas. We chose Texas as a place of operation since Dell's founder began the company in a Texas University dorm. The dataset we will be using is Austin workforce demographic data. The project's goal is also not new to Dell as they have diversity as one of their goals on their website.

Business Understanding

Business Objectives

- Increase Dell Technologies' (US) Black/African American and Hispanic/Latino presence within its workforce.
- Locate and focus on the Black/African American and Hispanic/Latino population with workforce potential.

Business Success Criteria

By 2030, 25% of Dell Technologies (US) workforce and 15% of its leaders identify as Black/African American and Hispanic/Latino minorities.

Assess the Situation

Risks

Lack of coverage is a massive risk for the datasets we will be analyzing. Realistically, no single dataset will give a complete picture of a city's workforce.

Condition

When covering an outlier, the contingency plan is to list all possible reasons why this may be the case from researching further the methods used to collect the data to cross-examining other similar data.

Contingency Plan

When covering a data value that is to be considered an outlier, the contingency plan is to list all possible reasons why this may be the case from researching further the methods used to collect the data to cross-examining other similar data.

Data Analysis Goals

Description

We'll analyze Austin, Texas's workforce based on ethnicity, job title, and age.

Dependency

This analysis will also study African American/Black and Hispanic/Latino potential employment within specific locations of Austin, Texas.

Prediction

This analysis will help determine the best locations to offer employment opportunities to African Americans/Blacks and Hispanics/Latinos communities.

Data Understanding

Inclusion/Exclusion Criteria

Figure 0.1

Variable Name	Included/Excluded	Rational
FISCAL YEAR	Included	
DEPARTMENT	Included	Department categorical attribute .
POSITION TITLE	Included	Position Title categorical value.
STAFFING LEVEL	Included	Staffing Level categorical value.
EMPLOYEE ETHNICITY	Included	Employee Ethnicity categorical value.
GENDER	Included	Gender categorical value.
GENERATION TYPE	Included	Generation Type categorical value.
EMPLOYEE CLASS ID	Included	Employee Class ID categorical value.
EMPLOYEE CLASS	Included	Employee Class categorical value.
EEOC CLASS ID	Included	EEOC Class ID categorical value
EEOC CLASS DESC	Included	Eeoc Class Desc categorical value.
AGE GROUP	Included	Age Group numerical value.
EFFECTIVE DATE	Included	
CITY	Included	City categorical value.
COUNTY CODE	Included	County Code categorical value.
COUNTY	Included	County categorical value.
STATE	Included	State categorical value.
ZIP CODE	Included	Zip Code categorical value.
EMPLOYEE COUNT	Included	Employee Count numerical value.
AVERAGE AGE	Included	Average Age numerical value.
AVG HOURLY RATE	Included	Avg Hourly Rate numerical value.

DEPARTMENTS

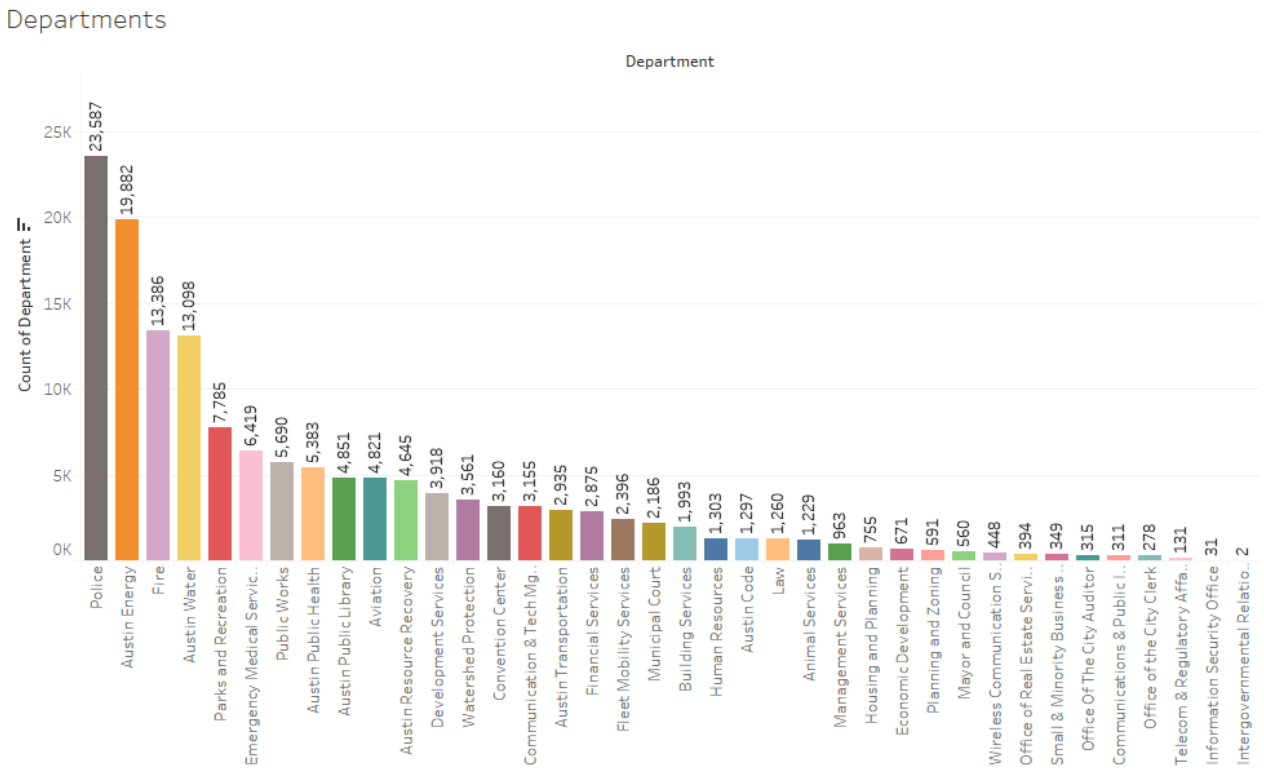
Describe Data

Figure 1.1

Categorical	DEPARTMENTS	COUNT	FREQUENCY
Attribute/Variable Name DEPARTMENTS	Police	23523	16%
	Austin Energy	19873	14%
	Fire	13370	9%
Data Volume (number of observation/rows) 146614	Austin Water	13100	9%
	Parks and Recreation	7785	5%
	Emergency Medical Services	6428	4%
Meaning of the attribute Departments of which individuals are employed under.	Public Works	5689	4%
	Austin Public Health	5384	4%
	Austin Public Library	4857	3%
	Aviation	4817	3%
Meaning of the attribute in business terms	Austin Resource Recovery	4641	3%
	Development Services	3917	3%
	Watershed Protection	3557	2%
	Convention Center	3158	2%
Attribute types (select from the list) Categorical	Communication & Tech Mgmt	3157	2%
	Austin Transportation	2928	2%
	Financial Services	2881	2%
	Fleet Mobility Services	2393	2%
	Municipal Court	2187	1%
	Building Services	1991	1%
	Human Resources	1305	1%
	Austin Code	1298	1%
	Law	1259	1%
	Animal Services	1230	1%
	Management Services	967	1%
	Housing and Planning	754	1%
	Economic Development	670	0%
	Planning and Zoning	591	0%
	Mayor and Council	560	0%
	Wireless Communication Svcs	448	0%
	Office of Real Estate Services	395	0%
	Small & Minority Business Rsrc	349	0%
	Office Of The City Auditor	318	0%
	Communications & Public Inform	309	0%
	Office of the City Clerk	279	0%
	Telecom & Regulatory Affairs	131	0%
	Information Security Office	31	0%
	Intergovernmental Relations	2	0%

Explore Data

Figure 1.2



This DEPARTMENT variable shows high counts for Police and Austin Energy. These having high counts are nothing unusual as Texas provides a lot of employment opportunities for law enforcement and energy employment.

Verify Data Quality

Figure 1.3

Variable Name	Data Quality Issue	Description/ Example	Problem
DEPARTMENTS	**Check coverage**	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	Yes
	Missing Attribute or blank fields	How will you address this?	No
	Duplicate	Duplicated records (observations)	No
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No
	Plausibility	E.g., all fields having the same or nearly the same values	No
	Conflict with Common Sense	E.g., teenagers with high income levels.	No
	High Cardinality	A high number of values in a set	No
	Outliers	An observation that lies well outside of the norm.	No
	Redundant Input	Does not give any new information that was not already explained by other inputs	No
	Sparseness	Any data which as very large zero value and very little no zero value	No
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No
	Unstructured Data	Unstructured data is data that does not follow a specified format	No

STAFFING LEVLES

Describe Data

Figure 2.1

Categorical	STAFFING LEVEL	COUNT	FREQUENCY
Attribute/Variable Name			
STAFFING LEVEL	Worker	116638	79.55%
	Supervisor	15499	10.57%
Data Volume (number of observation/rows)	Manager	11581	7.90%
	Executive	2144	1.46%
146614	Council Appointed	752	0.51%
Meaning of the attribute			
Level of employment status.			
Meaning of the attribute in business terms			
Workers report to supervisors. They report to managers, who report to executives.			
Attribute types (select from the list)			
Categorical			

Explore Data

Figure 2.2

This bar chart displays the expected ratio of workers to supervisors to managers to executive and lastly to the council appointed. The large count for subgroup workers is normal and nothing to be considered as an anomaly.

Verify Data Quality

Figure 2.3

Variable Name	Data Quality Issue	Description/ Example	Problem	Assessment
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
STAFFING LEVELS	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	Yes	Large percentage are workers.
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which has very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

EMPLOYEE ETHNICITY

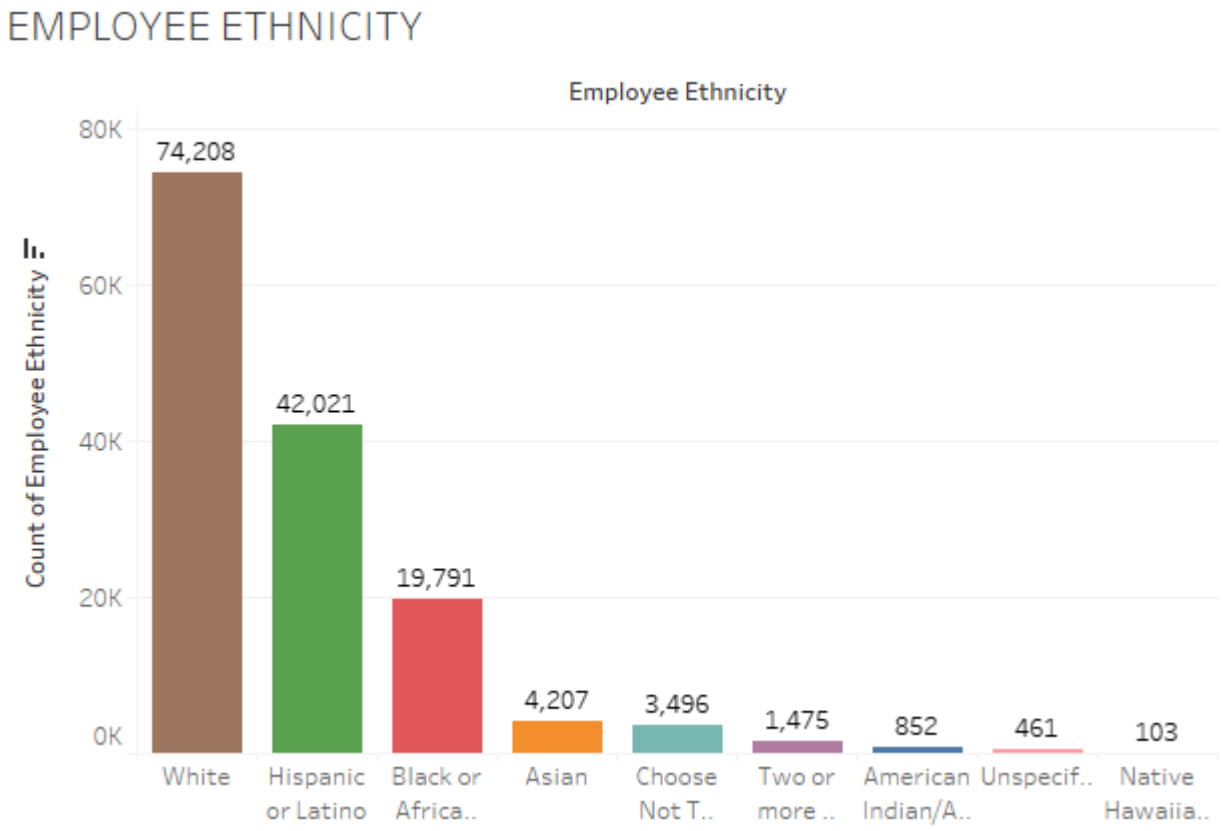
Describe Data

Figure 3.1

Categorical	EMPLOYEE ETHNICITY	Count	Percent
Attribute/Variable Name	White	74208	50.61%
EMPLOYEE ETHNICITY	Hispanic or Latino	42021	28.66%
	Black or African American	19791	13.50%
Data Volume (number of observation/rows)	Asian	4207	2.87%
146614	Choose Not to Disclose	3496	2.38%
	Two or more races	1475	1.01%
Meaning of the attribute	American Indian/Alaska Native	852	0.58%
Ethnic background of employees.	Unspecified	461	0.31%
	Native Hawaiian/Pacific Isl	103	0.07%
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			

Explore Data

Figure 3.2



This bar chart displays many White, Hispanic/Latinx, and Black/African. These large values are normal, and this chart does not show any abnormalities. It is well known in Texas to have a largely white majority.

Verify Data Quality

Figure 3.3

Variable Name	Data Quality Issue	Description/ Example	Problem	Assessment
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
EMPLOYEE ETHNICITY	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	Yes	Large value for white column.
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which has very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

GENDERS

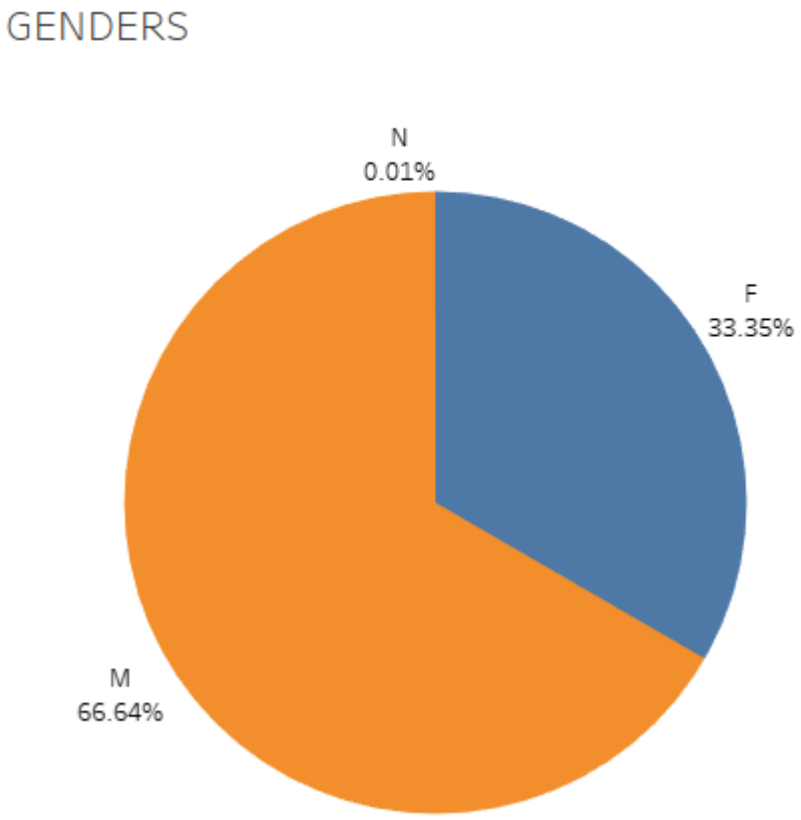
Describe Data

Figure 4.1

Categorical	GENDER	COUNT	FREQUENCY
Attribute/Variable Name	F	48901	33.35%
GENDER	M	97697	66.64%
	N	16	0.01%
Data Volume (number of observation/rows)			
14461			
Meaning of the attribute			
Biological association. Male, female, or n for non-disclosure.			
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			

Explore Data

Figure 4.2



This pie chart displays the ratio of male, female, and non-disclosed employees. Despite having 66.64% male and 33.35% female, what this data shows is that men have a significantly higher presence in the workforce demographic.

Verify Data Quality

Figure 4.3

Variable Name	Data Quality Issue	Description/ Example	Problem	Assessment
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
GENDERS	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which has very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

GENERATION TYPE

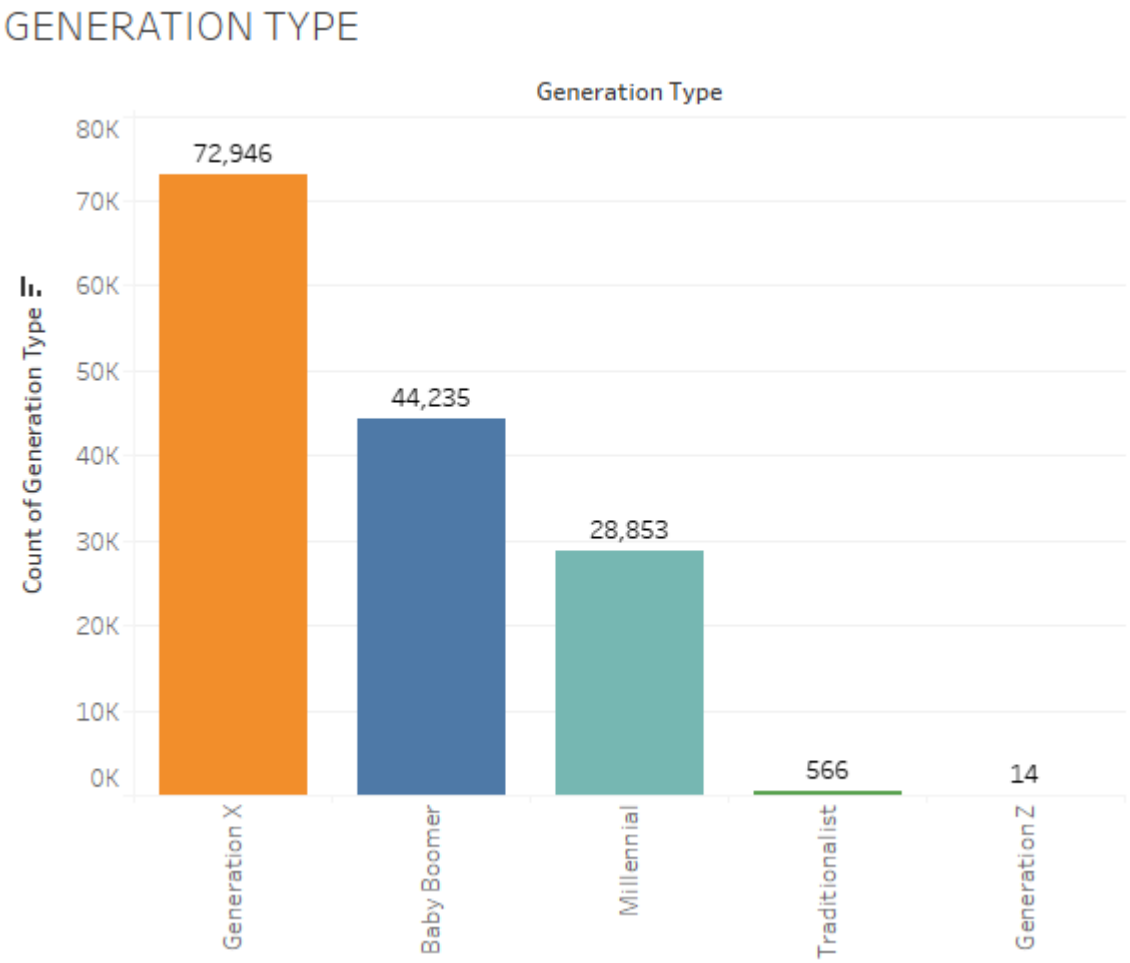
Describe Data

Figure 5.1

Categorical		Count	Percent
Attribute/Variable Name	Baby Boomer	44248	30%
GENERATION TYPE	Generation X	72922	50%
	Generation Z	13	0%
Data Volume (number of observation/rows)	Millennial	28783	20%
5	Traditionalist	566	0%
Meaning of the attribute			
Generation type is TX.			
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			

Explore Data

Figure 5.2



In this bar chart, Generation X shows a total of 72,946 records. Following as runner up is the Baby Boomers with 44,235 total records. These graph shows that there is many employees in Texas that are in the Generation X category.

Verify Data Quality

Figure 5.3

Variable Name	Data Quality Issue	Description/ Example	Problem
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No
GENERATION TYPE	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No
	Missing Attribute or blank fields	How will you address this?	No
	Duplicate	Duplicated records (observations)	No
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	N/A
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No
	Plausibility	E.g., all fields having the same or nearly the same values	No
	Conflict with Common Sense	E.g., teenagers with high income levels.	No
	High Cardinality	A high number of values in a set	No
	Outliers	An observation that lies well outside of the norm.	No
	Redundant Input	Does not give any new information that was not already explained by other inputs	No
	Sparseness	Any data which as very large zero value and very little no zero value	No
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No
	Unstructured Data	Unstructured data is data that does not follow a specified format	No

CITY

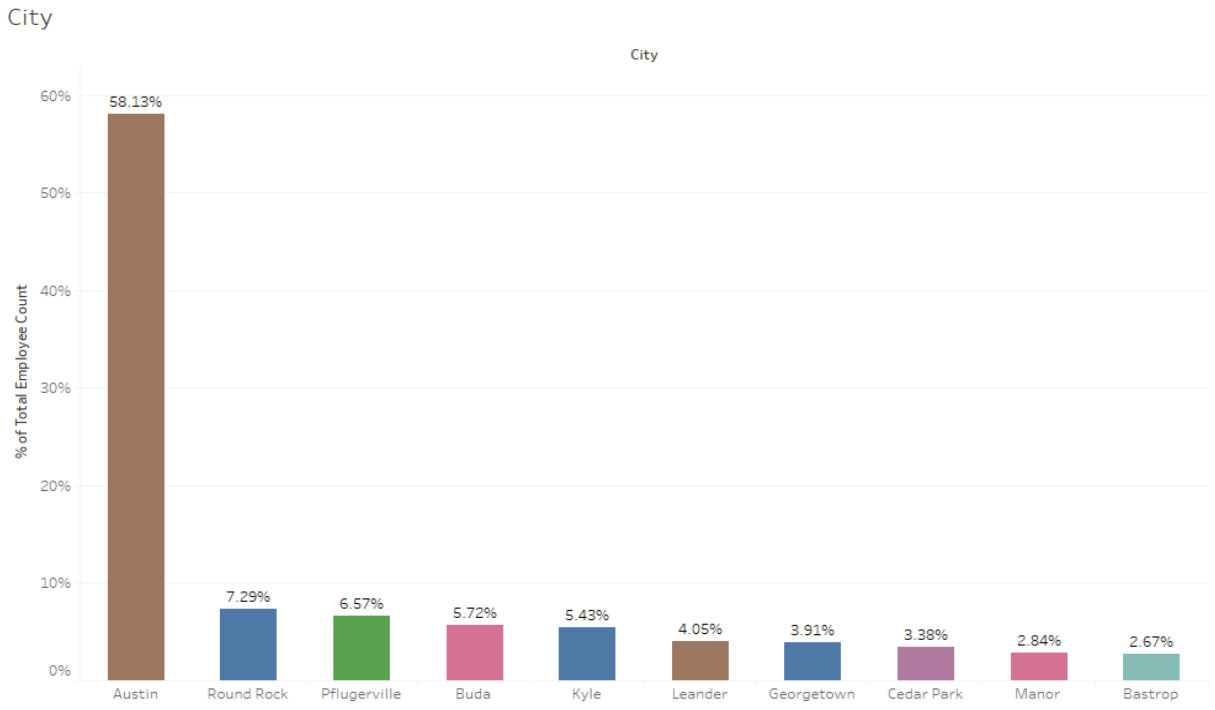
Describe Data

Figure 11.1

Categorical	Categories	Count	Percent
Attribute/Variable Name	Austin	62742	43%
City	Pflugerville	7094	5%
	Round Rock	7871	5%
Data Volume (number of observation/rows)	Buda	6175	4%
146533	Kyle	5857	4%
	Leander	4372	3%
Meaning of the attribute	Georgetown	4225	3%
	Cedar Park	3648	3%
Cities in Texas	Manor	3064	2%
	Bastrop	2884	2%
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			
**Excel: COUNTIF			

Explore Data

Figure 11.2



This bar chart depicts the difference in the volume of employees between the top five most populated cities. Austin having the largest volume was expected, since it is the capital of Texas. No abnormalities were found.

Verify Data Quality

Figure 11.3

Variable Name	Data Quality Issu	Description/ Example	Problem (Select fro	Assessment
City	Check coverage	represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attribute	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Commo	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	information about the target (dependent	No	
	Unstructured Data	Unstructured data is data that does not follow a	No	

EEOC CLASS DESCRIPTION

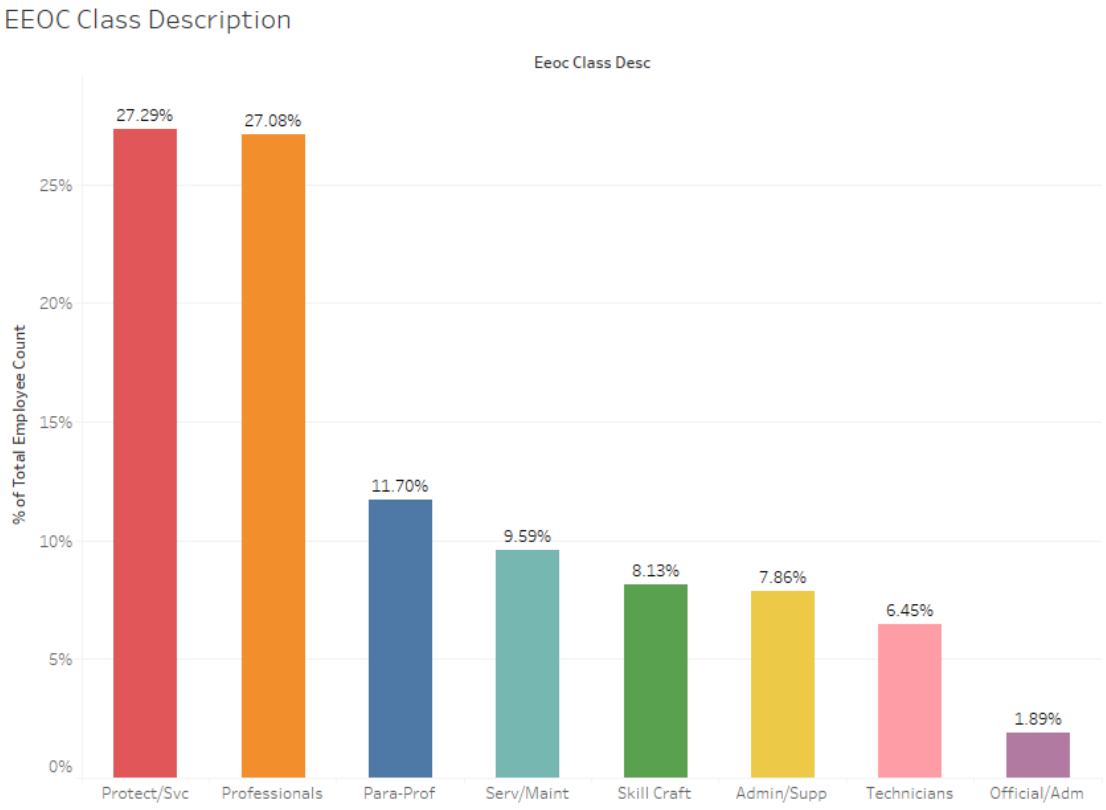
Describe Data

Figure 12.1

Categorical	Catergories	Count	Percent
Attribute/Variable Name	Professionals	39148	27%
EEOC Class DESCRIPTIONS	Serv/Maint	13865	10%
	Protect/Svc	39450	27%
Data Volume (number of observation/rows)	Para-Prof	16921	11%
146532	Skill Craft	11755	8%
	Admin/Supp	11367	8%
Meaning of the attribute	Technicians	9330	6%
	Official/Adm	2734	2%
Equal Employment Opportunity Commission Classes			
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			
**Excel: COUNTIF			

Explore Data

Figure 12.2



This bar chart depicts the volume of employees that are registered under their respective Equal Employment Opportunity Commission (EEOC) class. It was surprising that the Service/Maintenance was the fourth highest amongst the top five classes. Though, given the data, the result is valid. No abnormalities found.

Verify Data Quality

Figure 12.3

Variable Name	Data Quality Issu	Description/ Example	Problem (Select fr	Assessment
EEOC Class Descripti	Check coverage	represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attribute	of attributes and contained values fit together	No	
	Missing Attribute or I	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Commo	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

AGE GROUP

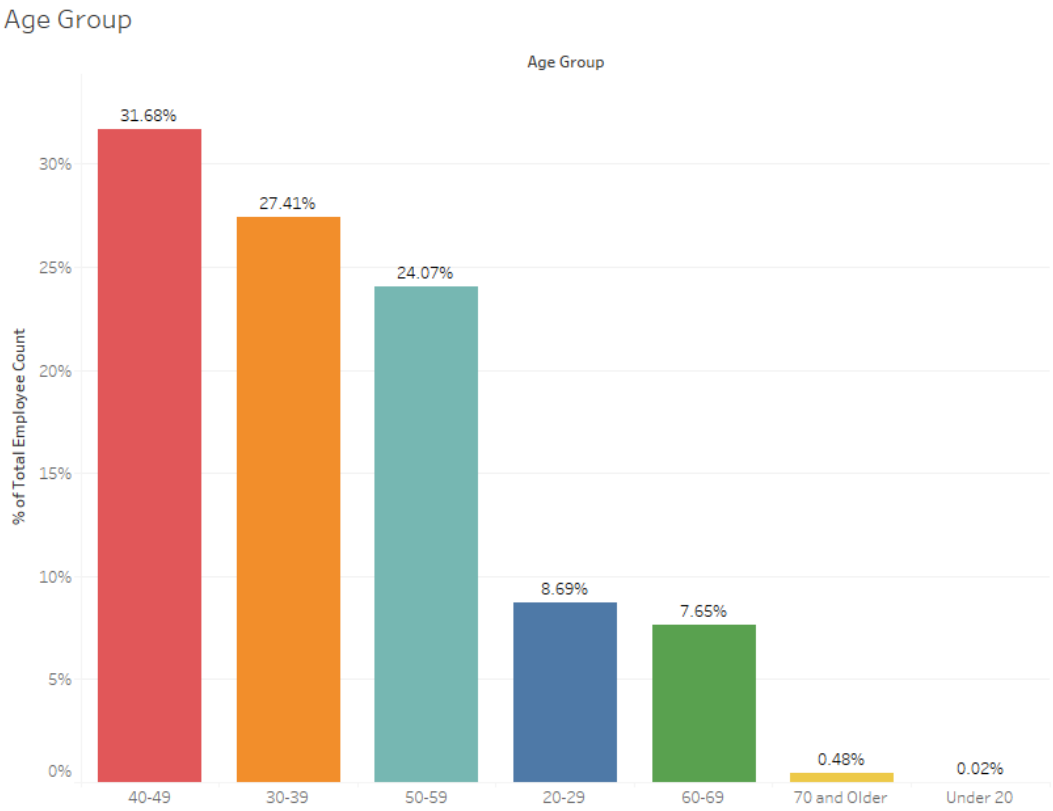
Describe Data

Figure 13.1

Categorical	Categories	Count	Percent
Attribute/Variable Name	40-49	45807	32%
Age Group	30-39	39623	28%
	50-59	34805	24%
Data Volume (number of observation/rows)	20-29	12560	9%
146533	60-69	11064	8%
Meaning of the attribute			
The Age Range of Employees			
Meaning of the attribute in business terms			
Attribute types (select from the list)			
Categorical			
**Excel: COUNTIF			

Explore Data

Figure 13.2



This bar chart depicts the number of employees is registered under a specific age group. The drop off at ages 40-49, 30-39, and 50-59 between 20-29 and 60-69 is quite common. No abnormalities found.

Verify Data Quality

Figure 13.3

Variable Name	Data Quality Issue	Description/ Example	Problem (Sele	Assessment
		represented •Use metadata (e.g., domain, range, dependency,		
Age Group	Check coverage	distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blar	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common S	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

EMPLOYEE CLASS

Describe Data

Figure 14.1

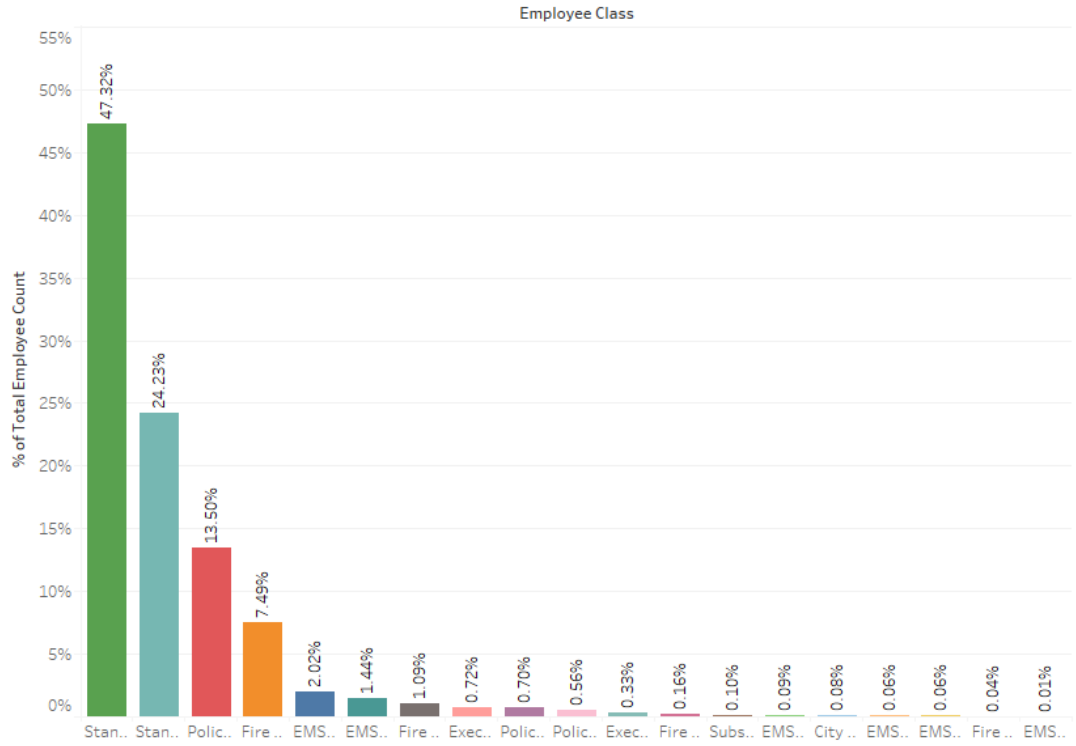
Categorical	Categories	Count	Percent
Attribute/Variable Name	Standard/Non-Exempt	68410	47%
Employee Class	Standard/Exempt	35033	24%
	Police/Non-Exempt	19514	14%
Data Volume (number of observation/rows)	Fire Kelly	10828	7%
146533	EMS42/Non-Exempt	2915	2%
	EMS48/Non-Exempt	2084	1%
Meaning of the attribute	Fire 40	1582	1%
	Executive/2	1035	1%
Classes of Employees	Police/Non-Exempt	1012	1%
	Police Cadet	804	1%
Meaning of the attribute in business terms	Executive/1	478	0%
	Fire Cadet	235	0%
	Substitute Judge Exempt	145	0%
	EMS40/Non-Exempt	130	0%
Attribute types (select from the list)	City Councilmembers	112	0%
Categorical	EMS40/Exempt	93	0%
	EMS48/Exepmt	87	0%
	Fire 40 Exempt	58	0%
**Excel: COUNTIF	EMS42/Exempt	15	0%

** All data after EMS42/Exempt are less than 0.00%*

Explore Data

Figure 14.2

Employee Class



This bar chart depicts the volume of employees based on their respective employee class. The results were within expectations. No abnormalities found.

Verify Data Quality

Figure 14.3

Variable Name	Data Quality Issue	Description/ Example	Problem (Select fro	Assessment
Employee Class	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	N/A	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fie	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	N/A	
	Plausibility	same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	N/A	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

EMPLOYEE CLASS ID

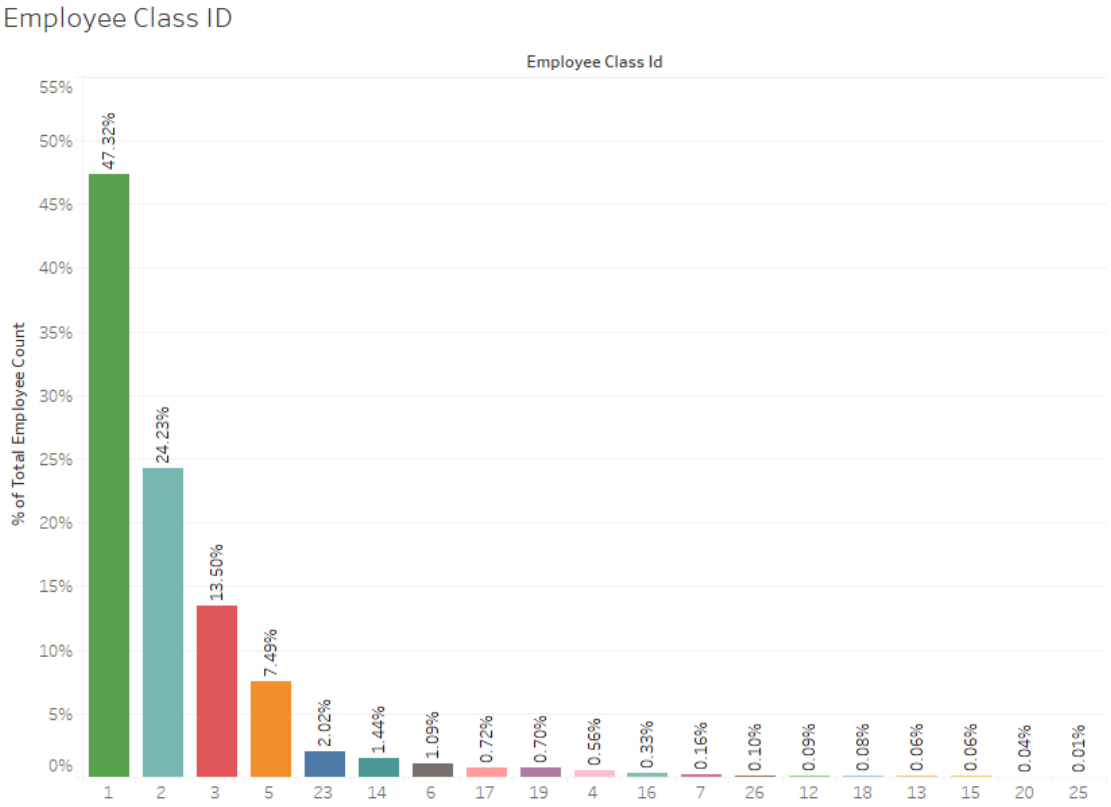
Describe Data

Figure 15.1

Categorical	Categories	Count	Percent
Attribute/Variable Name	Class 1	68410	47%
Employee Class ID	Class 2	35033	24%
	Class 3	19514	14%
Data Volume (number of observation/rows)	Class 4	804	1%
146533	Class 5	10828	7%
	Class 6	1582	1%
Meaning of the attribute	Class 7	235	0.20%
	Class 12	130	0.10%
Employee Class Indentification	Class 13	93	0.10%
	Class 14	2084	1%
Meaning of the attribute in business terms	Class 15	87	0.10%
	Class 16	478	0.30%
	Class 17	1035	1.00%
	Class 18	112	0.10%
Attribute types (select from the list)	Class 19	1012	1%
Categorical	Class 20	58	0%
	Class 23	2915	2%
	Class 25	15	0%
	Class 26	145	0.10%
**Excel: COUNTIF			

Explore Data

Figure 15.2



This bar chart depicts the volume of employees based on their employee class id. Since this chart reflects the results of the Employee Class bar chart, we can confirm that the results are valid. No abnormalities found.

Verify Data Quality

Figure 15.3

Variable Name	Data Quality Issue	Description/ Example	Problem (Select from	Assessment
Employee ID	Check coverage	represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blanl	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Se	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

COUNTY CODE

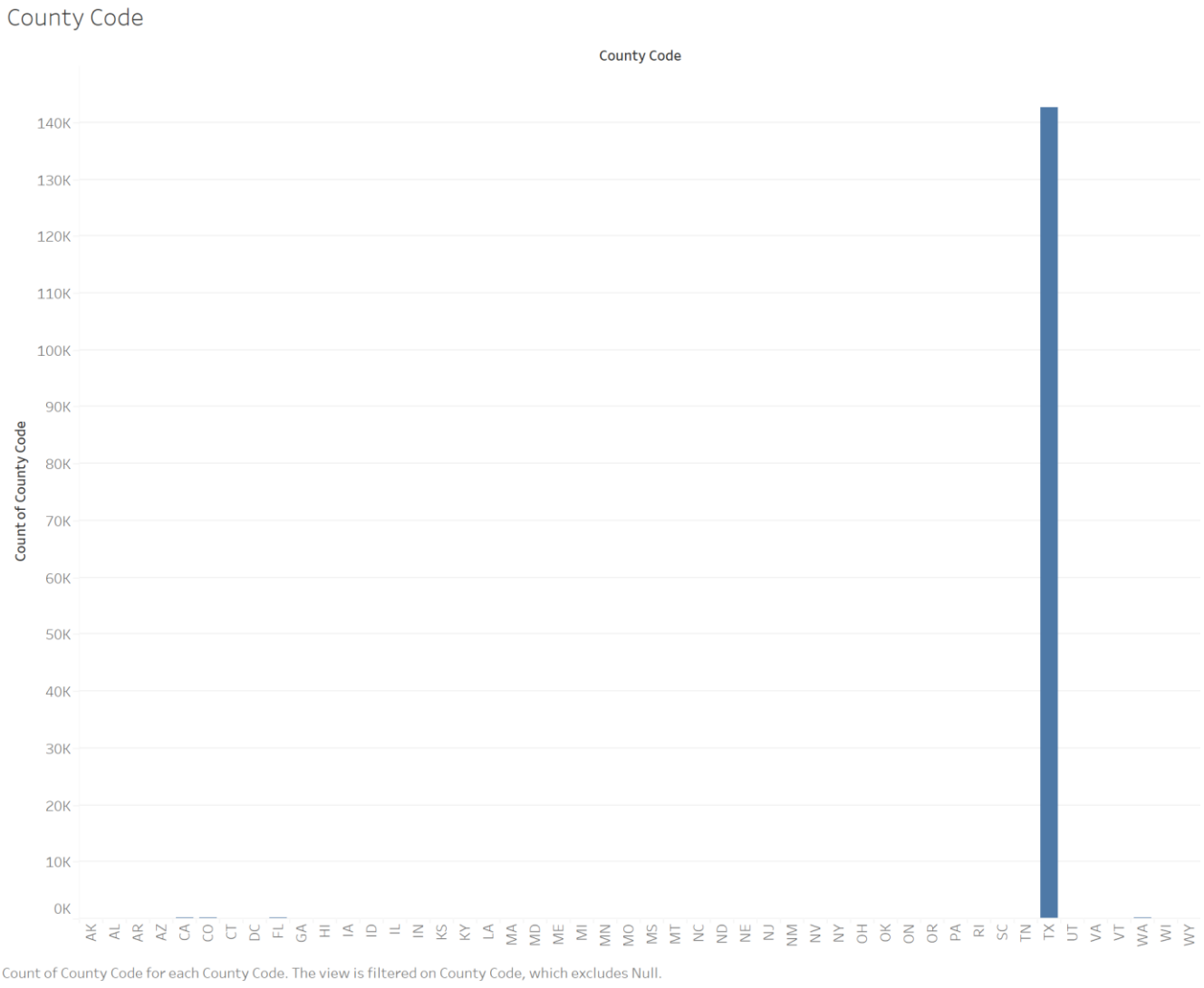
Describe Data

Figure 6.1

Categorical	Class Label (e.g. categories of earnings Hourly, weekly, biweekly...annually)	Count	Percentage
Attribute/Variable Name:	TX	142953	99.04%
County Code	CA	194	0.13%
	CO	136	0.09%
	WA	112	0.08%
	FL	110	0.08%
	VA	91	0.06%
	AZ	56	0.04%
	NC	54	0.04%
	IL	50	0.03%
	MI	45	0.03%
	PA	44	0.03%
Data Volume (number of observation/rows):	OK	44	0.03%
144544	NY	43	0.03%
	AL	39	0.03%
	NM	39	0.03%
Meaning of the attribute:	OR	38	0.03%
The county code of each employee within this	MA	37	0.03%
data set	TN	34	0.02%
	UT	33	0.02%
Attribute types (select from the list):	MN	32	0.02%
Categorical	WI	32	0.02%
	GA	26	0.02%
	LA	26	0.02%
	AR	26	0.02%
	OH	23	0.02%
	SC	22	0.02%

Explore Data

Figure 6.2



Although the title of figure #.# is "County Code," the data it contains depicts the state in which our data originates. According to the graph, 99% of our data comes from Texas, which coincides with this project's purpose.

Verify Data Quality

Figure 6.3

Variable Name	Data Quality Issue	Description/ Example	Problen	Assessment
County Code	**Check coverage**	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	The values are states not county code.
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	

COUNTY

Describe Data

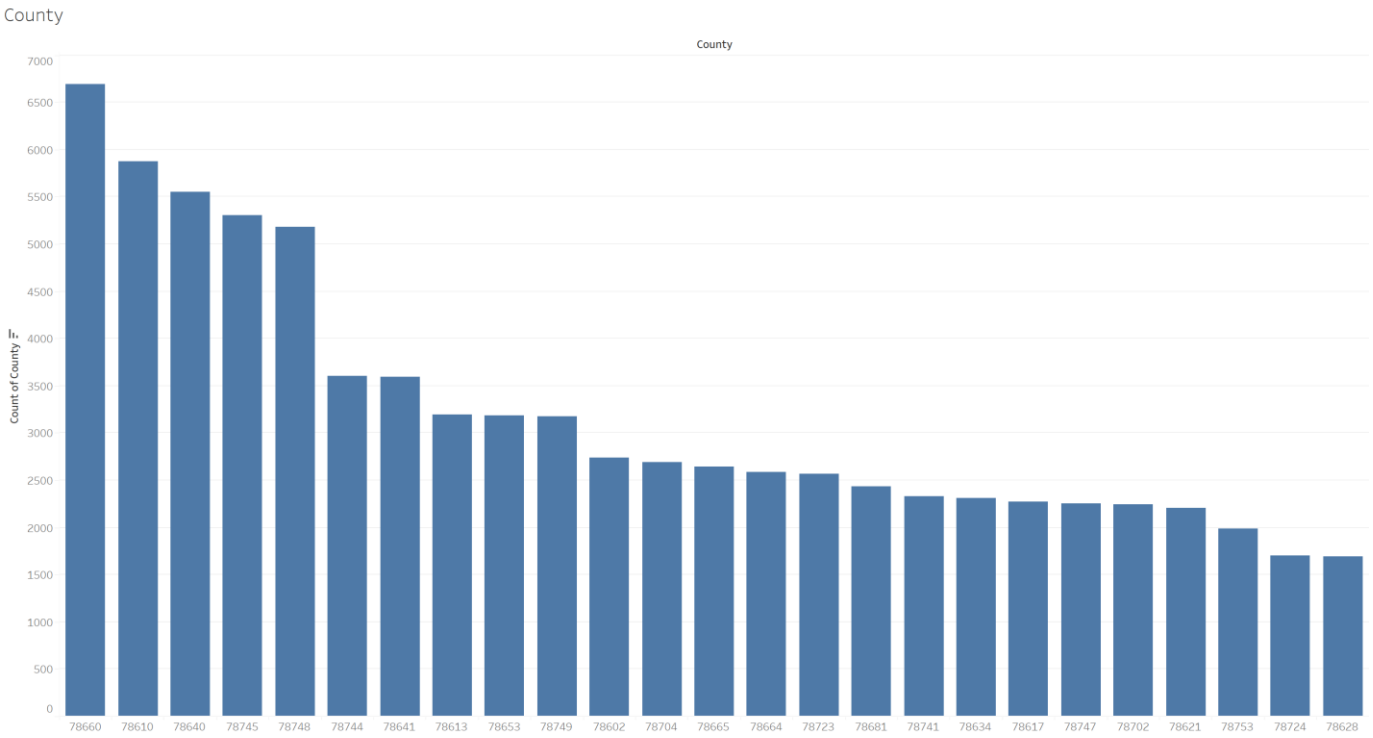
Figure 7.1

Categorical	Class Label (e.g. categories of earnings Hourly, weekly, biweekly...annually)	Count	Percentage
Attribute/Variable Name:	78660	6702	4.64%
County	78610	5773	3.99%
	78745	5565	3.85%
	78748	5493	3.80%
	78640	5280	3.65%
	78744	3795	2.63%
	78641	3476	2.40%
	78749	3297	2.28%
	78613	3272	2.26%
	78704	3056	2.11%
	78653	3017	2.09%
Data Volume (number of observation/row	78665	2714	1.88%
144547	78723	2677	1.85%
	78664	2656	1.84%
	78741	2586	1.79%
Meaning of the attribute:	78602	2545	1.76%
The county of each employee within this	78702	2463	1.70%
data set	78681	2412	1.67%
	78617	2312	1.60%
Attribute types (select from the list):	78634	2272	1.57%
Categorical	78747	2251	1.56%
	78753	2196	1.52%
	78621	2103	1.45%
	78754	1786	1.24%
	78724	1738	1.20%
	78758	1675	1.16%

**Data was cut short to preserve the length of this table. Data values after 4832 is less than 1.16%*

Explore Data

Figure 7.2



As with the issues of figure #.#, figure #.# also has a misleading name. Instead of depicting county names within this dataset, the graph "County" displays zip codes. Googling the most frequent zip codes, we find that majority of these zip codes reside in Austin, Texas.

Verify Data Quality

Figure 7.3

Variable Name	Data Quality Issue	Description/ Example	Problem (Se	Assessment
County	**Check coverage**	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fiel	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	The values are zip codes not county names.
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

STATE

Describe Data

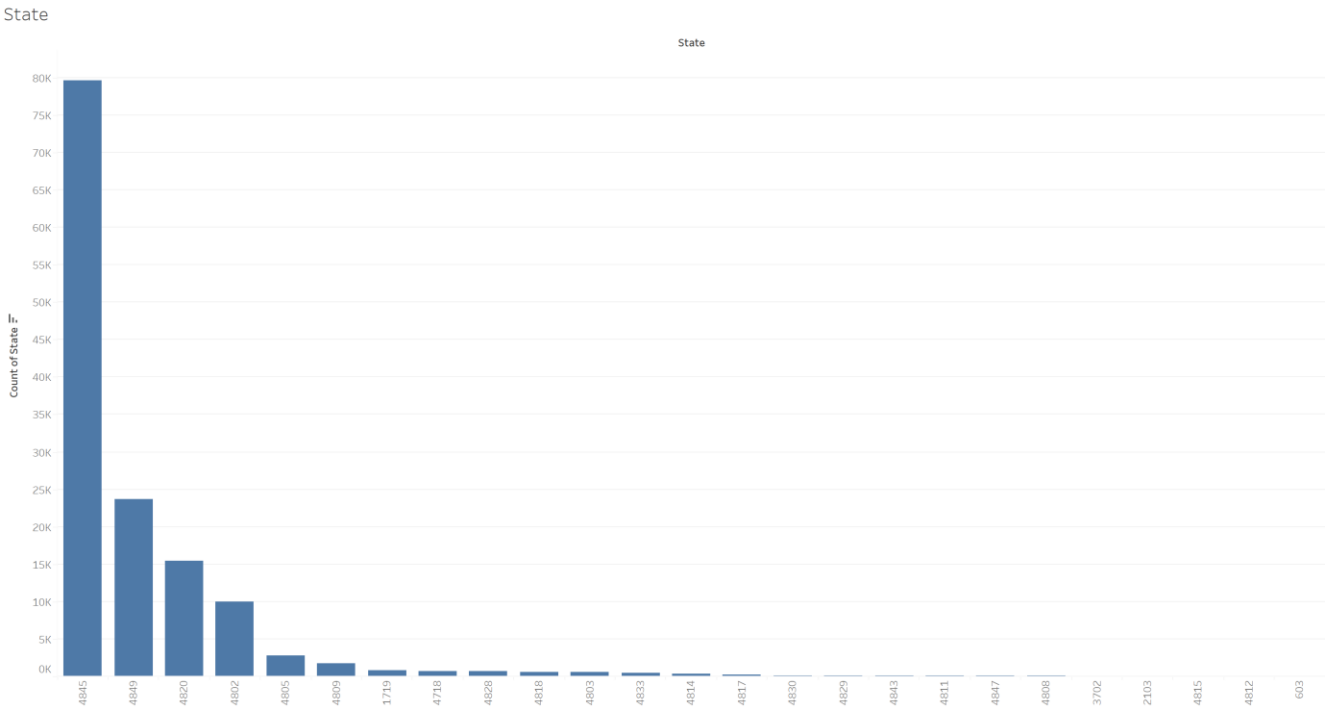
Figure 8.1

Categorical	Class Label (e.g. categories of earnings Hourly, weekly, biweekly...annually)	Count	Percentage
Attribute/Variable Name:	4845	81890	58.34%
State	4849	23284	16.59%
	4820	14702	10.47%
	4802	9588	6.83%
	4805	2809	2.00%
	4809	1463	1.04%
	1719	905	0.64%
	4718	748	0.53%
	4828	597	0.43%
	4818	592	0.42%
	4803	493	0.35%
Data Volume (number of observation/rows):	4833	348	0.25%
140370	4814	262	0.19%
	4817	193	0.14%
	4830	137	0.10%
Meaning of the attribute:	4811	97	0.07%
The state of each employee within this data set	4843	95	0.07%
	4829	89	0.06%
	4847	58	0.04%
Attribute types (select from the list):	2103	55	0.04%
Categorical	3108	47	0.03%
	4815	47	0.03%
	3702	46	0.03%
	4804	42	0.03%
	4808	40	0.03%
	4832	39	0.03%

**Data was cut short to preserve the length of this table. Data values after 4832 is less than 0.03%*

Explore Data

Figure 8.2



Following the trend, Figure #.# is mislabeled. Instead of depicting states as its title suggests, its data contain county codes. It didn't prove easy when verify the location of these county codes because different agencies use different county code systems. Thankfully, looking at the dataset as a whole, these county codes are attached to Austin, Texas.

Verify Data Quality

Figure 8.3

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment
State	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	The values contain county codes not states.
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

ZIP CODE

Describe Data

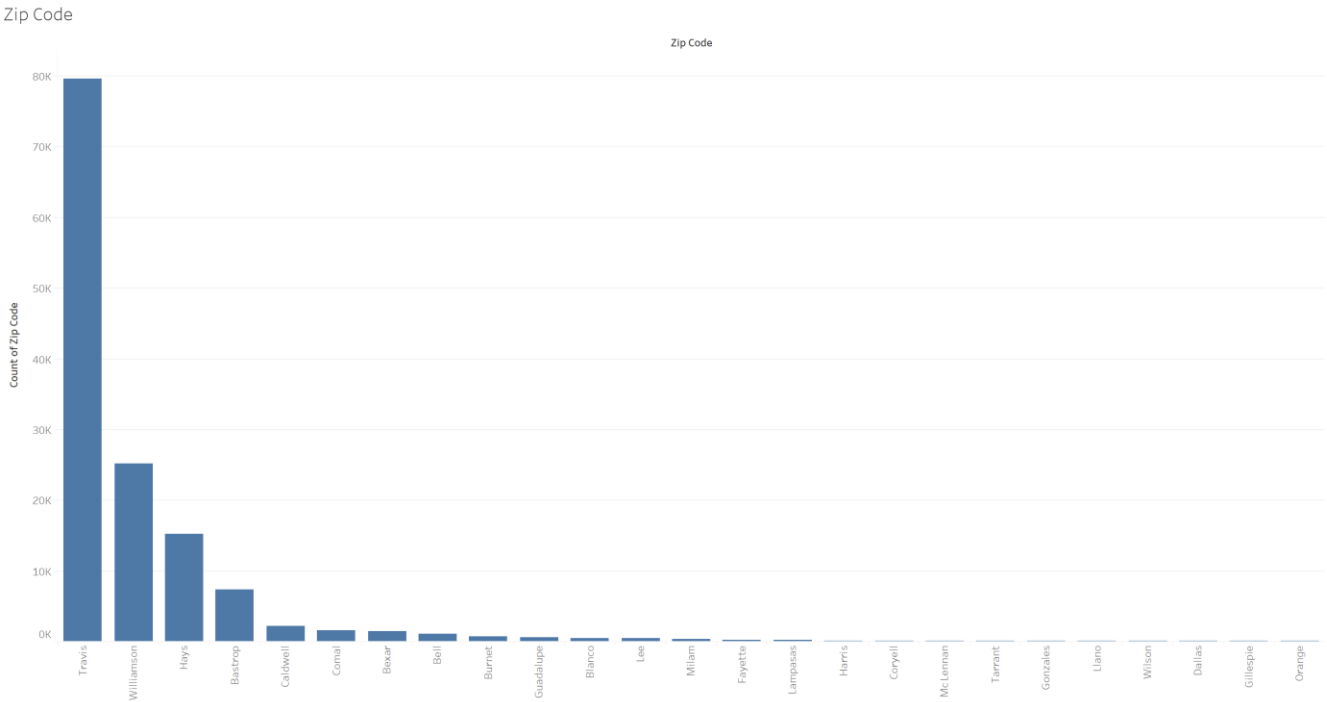
Figure 9.1

Categorical	Class Label (e.g. categories of earnings Hourly, weekly, biweekly...annually)	Count	Percentage
Attribute/Variable Name:	Travis	81880	58.33%
Zip Code	Williamson	24861	17.71%
	Hays	14571	10.38%
	Bastrop	7060	5.03%
	Caldwell	2244	1.60%
	Bexar	1467	1.05%
	Comal	1318	0.94%
	Bell	1069	0.76%
	Burnet	651	0.46%
	Guadalupe	600	0.43%
	Blanco	453	0.32%
Data Volume (number of observation/row	Lee	438	0.31%
140370	Milam	297	0.21%
	Fayette	238	0.17%
	Lampasas	171	0.12%
Meaning of the attribute:	Harris	140	0.10%
The zip code of each employee within the	Mc Lennan	135	0.10%
data set	Coryell	132	0.09%
	Gonzales	108	0.08%
Attribute types (select from the list):	Dallas	99	0.07%
Categorical	Tarrant	94	0.07%
	Wilson	89	0.06%
	Gillespie	85	0.06%
	Llano	83	0.06%
	Burleson	47	0.03%
	Fort Bend	47	0.03%

**Data was cut short to preserve the length of this table. Data values after 4832 is less than 0.03%*

Explore Data

Figure 9.2



Last of the mislabeled graphs is figure #.#. Instead of depicting zip codes, its value represents county names. The county names coincide with the largest county in Austin, Texas. This graph further proves the reliability of this dataset.

Verify Data Quality

Figure 9.3

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment
Zip Code	**Check coverage**	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	The values contain county names not zip codes.
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

EMPLOYEE COUNT

Describe Data

Figure 10.1

Categorical	Class Label (e.g. categories of earnings Hourly, weekly, biweekly…annually)	Count/Frequency	Percentage
Attribute/Variable Name:	1	140384	95.70%
Employee Count	2	4475	3.05%
	3	908	0.62%
	4	394	0.27%
	5	164	0.11%
	6	126	0.09%
	7	60	0.04%
	8	40	0.03%
	9	32	0.02%
	10	29	0.02%
	11	14	0.01%
Data Volume (number of observation/rows):	12	13	0.01%
146695	13	17	0.01%
	14	5	0.00%
	15	4	0.00%
Meaning of the attribute:	16	11	0.01%
Employee representation within the data set	17	8	0.01%
	18	6	0.00%
	20	3	0.00%
Attribute types (select from the list):	22	1	0.00%
Categorical	25	1	0.00%

Explore Data

Figure 10.2

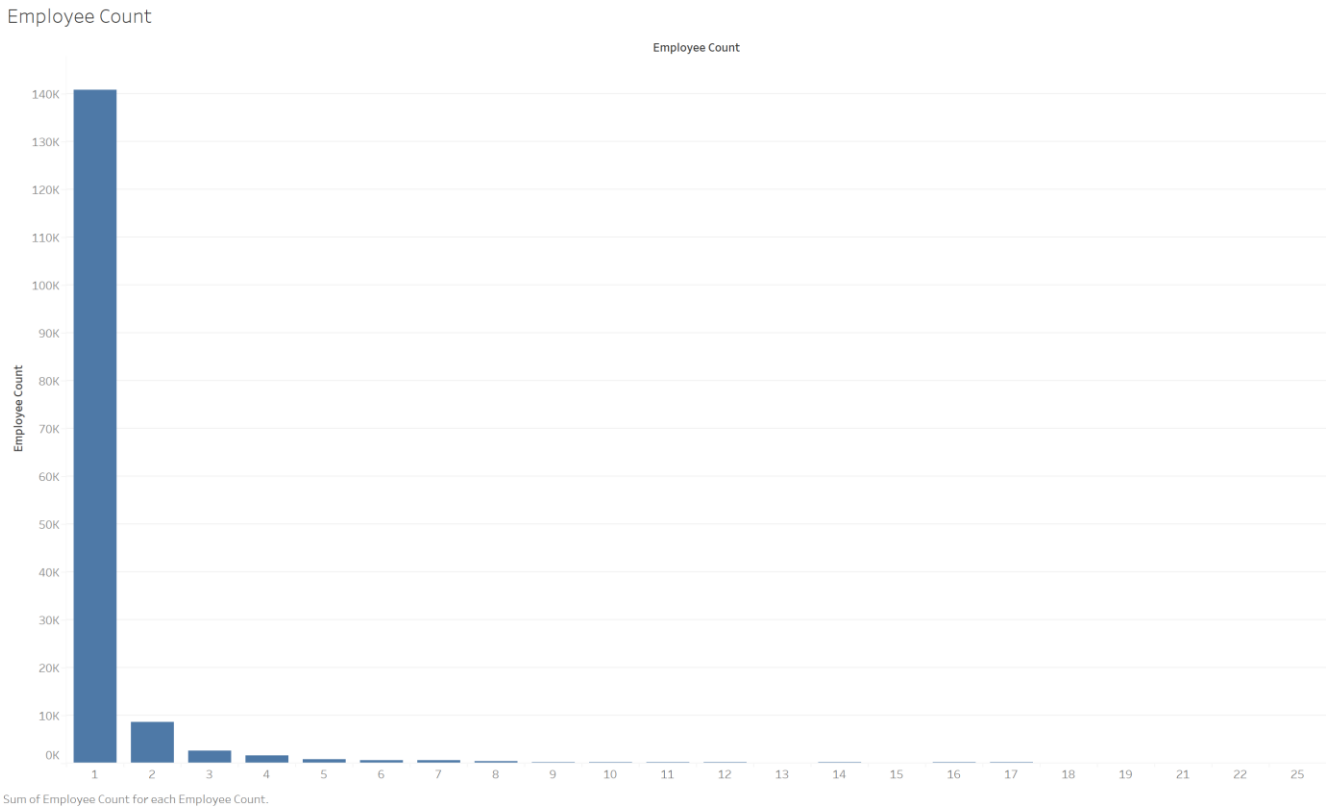


Figure 10.2 depicts the representation of the entire dataset. In order words, each variable is unique and represents a single employee. 4.3% of the data represent more than a single employee.

Verify Data Quality

Figure 10.3

Variable Name	Data Quality Issue	Description/ Example	Problem	Assessment
Employee Count	**Check coverage**	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No	
	Meaning Of Attributes	Verify that the meanings of attributes and contained values fit together	No	
	Missing Attribute or blank fields	How will you address this?	No	
	Duplicate	Duplicated records (observations)	No	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No	
	Plausibility	E.g., all fields having the same or nearly the same values	No	
	Conflict with Common Sense	E.g., teenagers with high income levels.	No	
	High Cardinality	A high number of values in a set	No	
	Outliers	An observation that lies well outside of the norm.	No	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No	
	Sparseness	Any data which as very large zero value and very little no zero value	No	
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No	

Data Preparation

County Code

Figure 6.1's title is "County Code"; however, its values do not represent county codes. Instead, it displays the states from where it originates. Figure 6.1's new title will be "States" to prevent further confusion.

County

Like figure 6.1 figure 7.1's title is "County"; however, its values do not represent county names. Instead, it displays the Zip Codes from which the data originates. Figure 7.1's new title will be "Zip Codes" to prevent further confusion.

State

Figure 8.1's title is "State"; however, its values do not represent states. Instead, it displays county codes from which the data originates. Figure 8.1's new title will be "County Codes."

Zip Code

Last but not least, figure 9.1's title is "Zip Codes"; however, its values do not represent zip codes. Instead, it displays the county names from which the data originates. Figure 9.1's new title will be "County."

During the creation of this data set, "States," "Zip Codes," "County Codes," and "Zip Code" was mislabeled. Our solution is to relabel them to suit their value.

Modeling

Figure 11.1

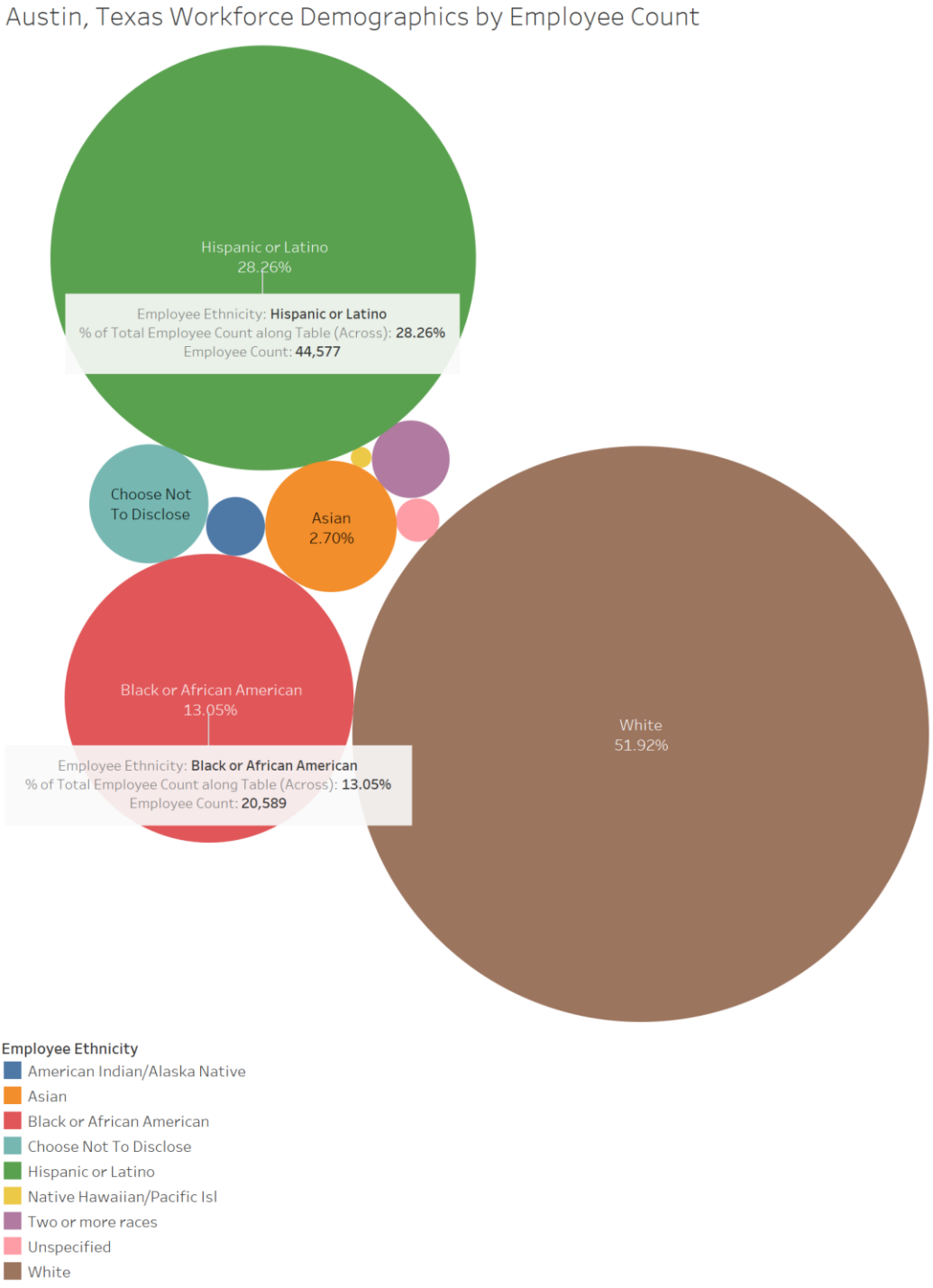


Figure 11.1 depicts Austin, Texas' workforce demographic by ethnicity. 51.92% of individuals identify as White, 28.26% identify as Hispanic or Latino, 13.05% identify as Black or African American, and 2.70% identify as Asian. 41.31% (combining Black or African American and Hispanic or Latino) of Austin, Texas identify as the ethnicity Dell is targeting.

Figure 11.2

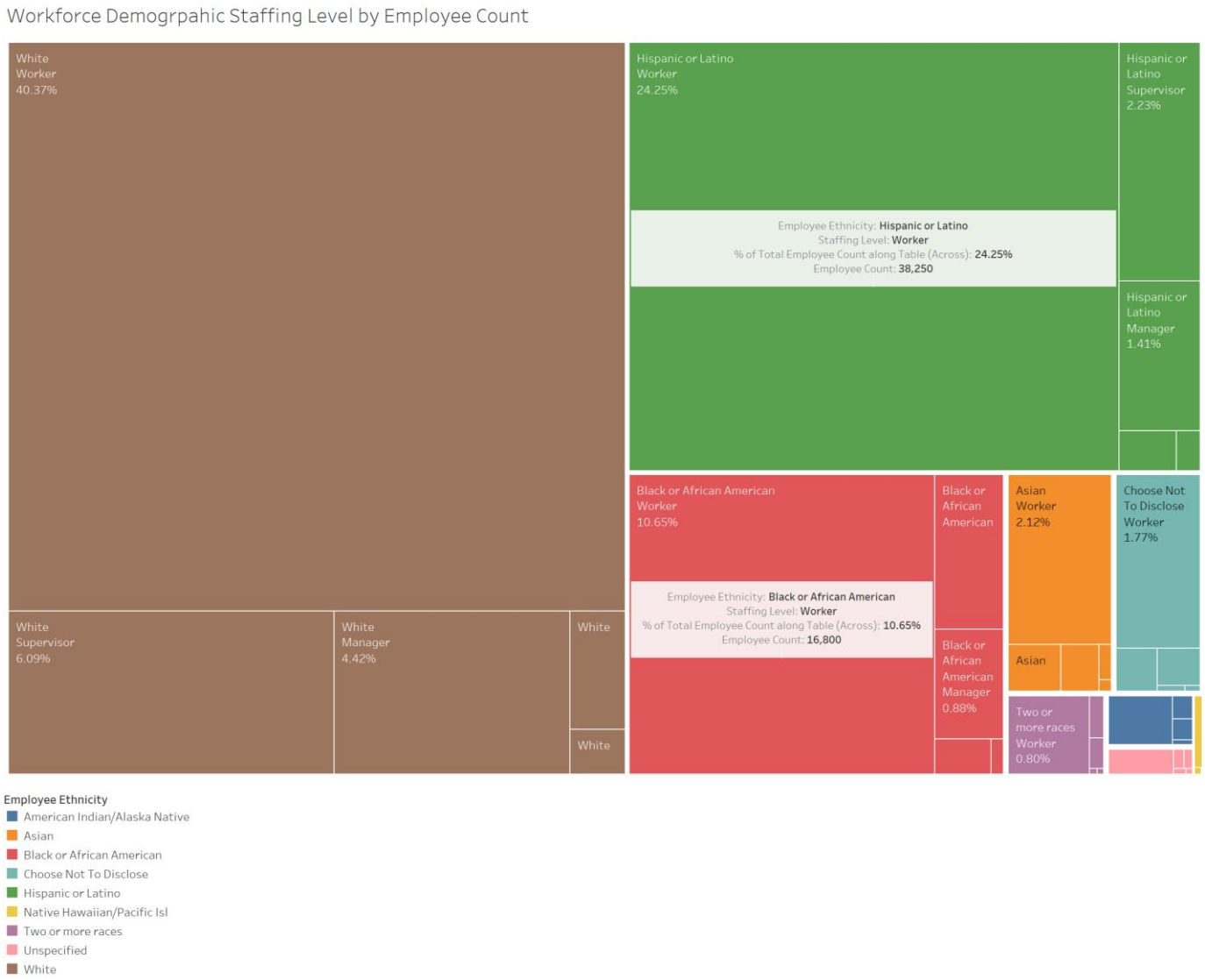


Figure 11.2 drills deeper into Austin, Texas' workforce demographic and displays the staffing level within an ethnic population. 24.25% of individuals who identify as Hispanic or Latino are workers. 2.23% of them are supervisors, and 1.41% are managers. 10.65% of individuals who identify as Black or African American are workers. 1.24% of them are supervisors, and 0.88% are managers.

Figure 11.3

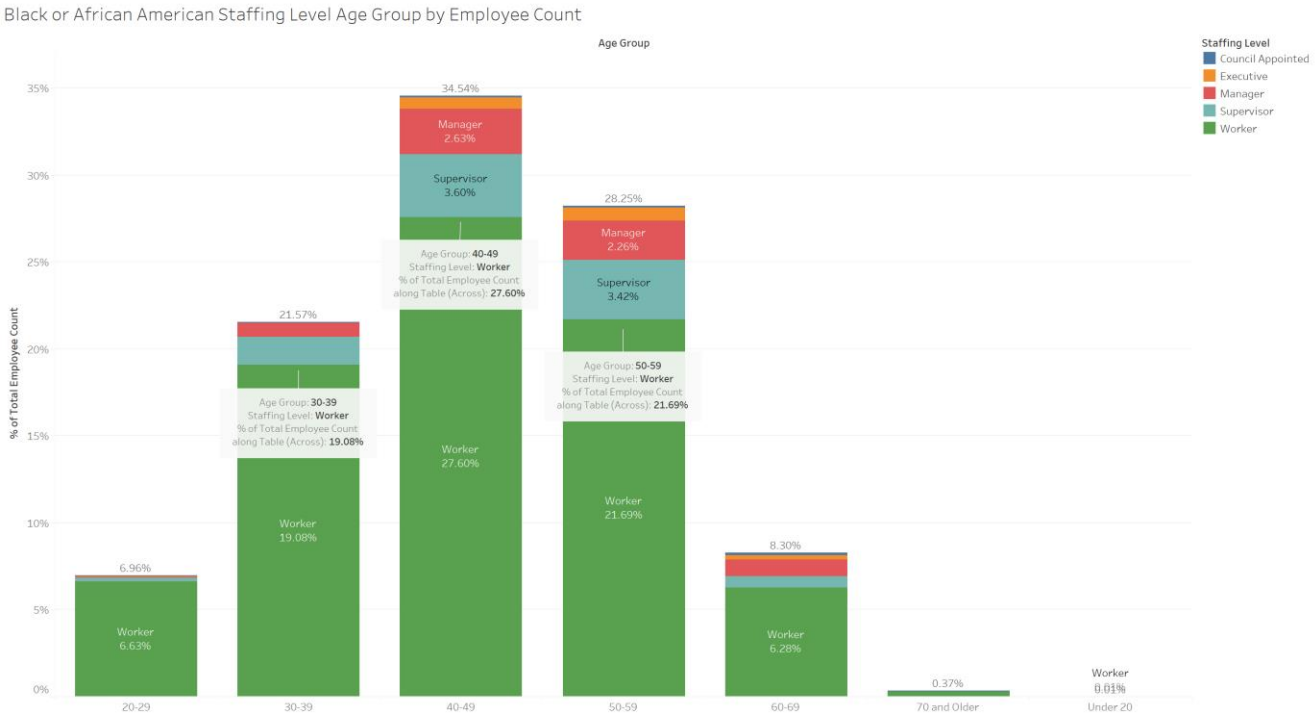


Figure 11.3 goes further and looks at Austin, Texas' Black or African American staffing level and groups them into their respective age group. It finds a considerable portion of workers are in the 30 to 39, 40 to 49, and 50 to 59 age groups. The percentage of supervisors and managers between 30 to 39 is less than half of the supervisors of 40 to 49 and 50 to 59.

Figure 11.4

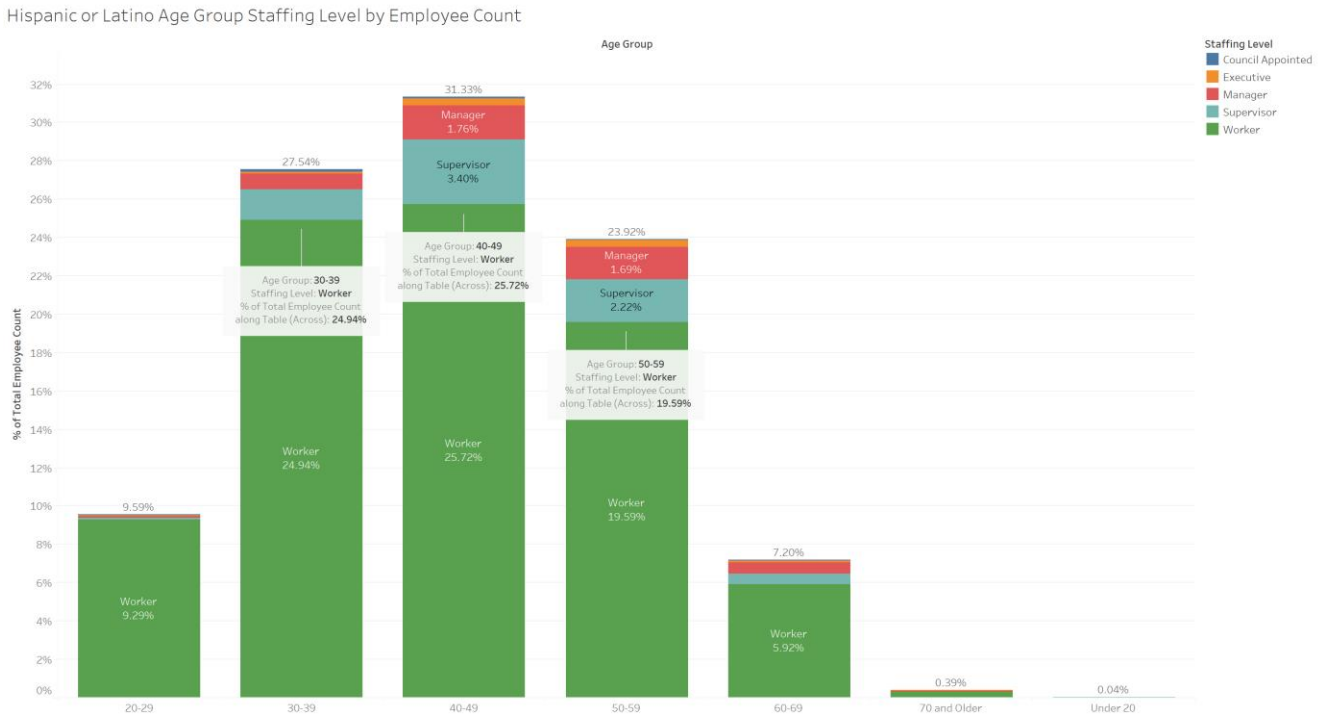


Figure 11.4 goes further and looks at Austin, Texas' Hispanic or Latino staffing level and groups them into their respective age group. It finds a considerable portion of workers are in the 30 to 39, 40 to 49, and 50 to 59 age groups. The percentage of supervisors and managers between 30 to 39 is less than half of the supervisors of 40 to 49 and 50 to 59.

Figure 11.5

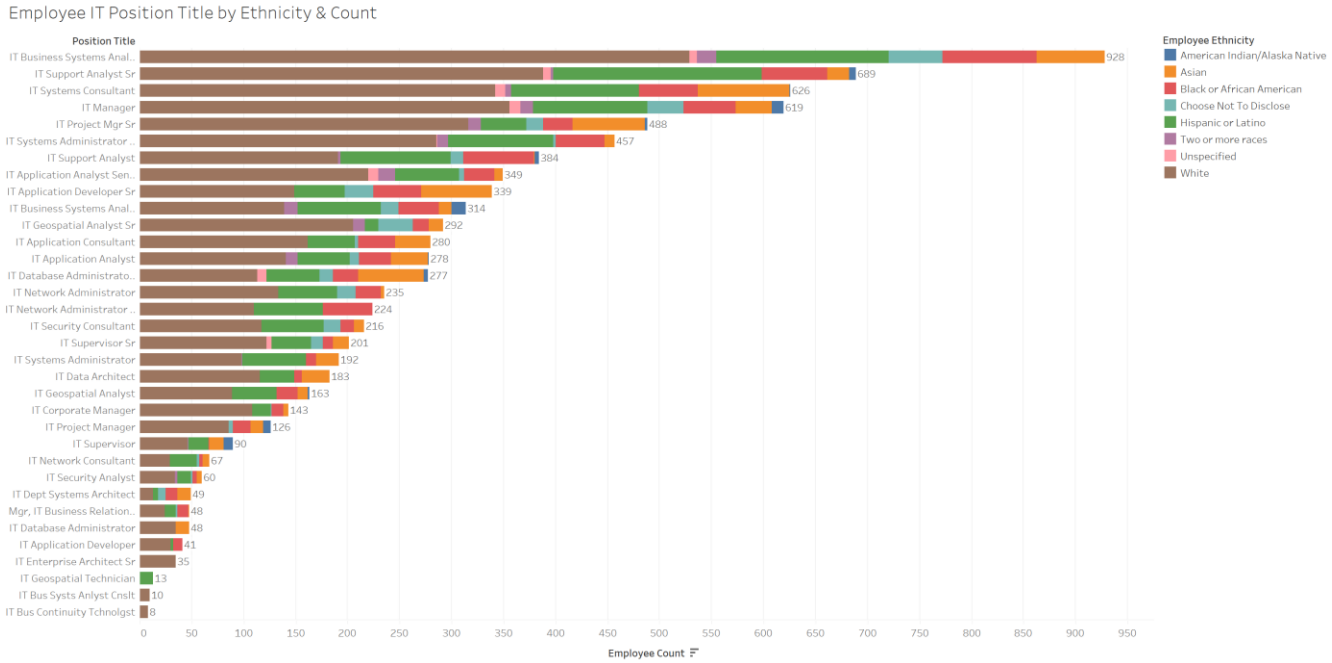


Figure 11.5 displays IT position titles by ethnicity. It finds positions such as IT project manager, IT supervisor, IT Network Consultant, IT Security Analyst, IT Dept Systems Architect, Manager of IT Business Relations, IT Database Administrator, IT Application Developer, IT Enterprise Architect, IT Geospatial Technician, IT Bus Analyst Consultant, and IT Business Continuity Technology lack representation of Black or African American and Hispanic or Latino representation.

Figure 11.6

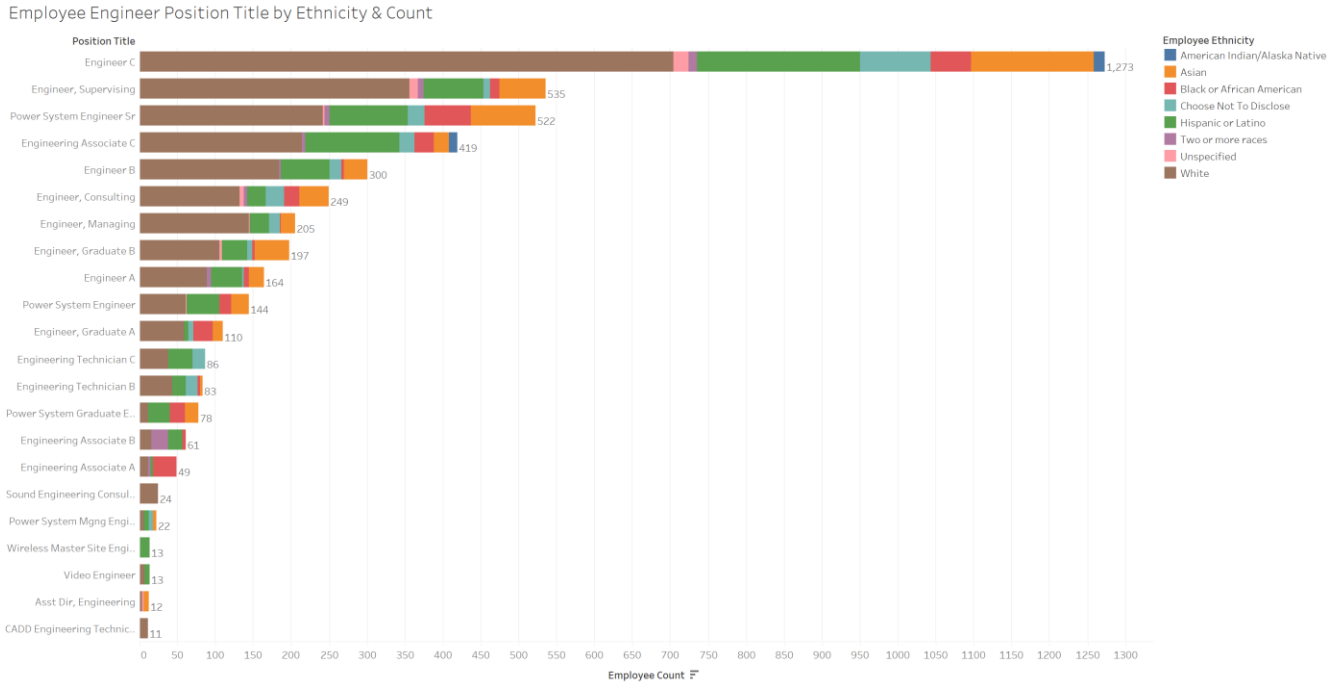


Figure 11.6 displays the Engineering position titles by ethnicity. It finds positions such as Engineering Associate A, Sound Engineering Consultant, Power System Management Engineer, Wireless Master Site Engineer, Video Engineer, Assistant Director of Engineering, and CADD Engineering Technician lack representation of Black or African American and Hispanic or Latino population.

Figure 11.7

Multiple Regression for AVG HOURLY RATE *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.2482	0.0616	0.0614	13.28082637	96763	0

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	8	579993.5285	72499.19107	411.0389364	< 0.0001
Unexplained	50102	8837008.247	176.380349		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	37.75974495	0.084103042	448.9700262	< 0.0001	37.59490203	37.92458786
EMPLOYEE ETHNICITY (American Indian/Alaska Native)	-3.189432447	0.83430154	-3.822877333	0.0001	-4.824672922	-1.554191972
EMPLOYEE ETHNICITY (Asian)	3.240255053	0.330192793	9.81322161	< 0.0001	2.593073436	3.88743667
EMPLOYEE ETHNICITY (Black or African American)	-5.792596062	0.183446556	-31.57647759	< 0.0001	-6.152153391	-5.433038733
EMPLOYEE ETHNICITY (Choose Not To Disclose)	-1.588670567	0.390950387	-4.0636117	< 0.0001	-2.354937757	-0.822403377
EMPLOYEE ETHNICITY (Hispanic or Latino)	-6.953280107	0.138300119	-50.27674693	< 0.0001	-7.224349909	-6.682210306
EMPLOYEE ETHNICITY (Native Hawaiian/Pacific Isl)	-11.36991444	1.731060589	-6.568178209	< 0.0001	-14.76281281	-7.977016063
EMPLOYEE ETHNICITY (Two or more races)	-4.460371906	0.532476806	-8.37665013	< 0.0001	-5.504032482	-3.416711331
EMPLOYEE ETHNICITY (Unspecified)	8.394101207	3.684398523	2.278282643	0.0227	1.172638341	15.61556407

Figure 11.7 is a regression table created using StatTools to test the relationship between Employee Ethnicity (independent) and Average Hourly Rate (dependent). We hypothesized that employee ethnicity directly affects an employee's average hourly rate. Although the R-Square is small, its p-value is less than 0.05, rejecting our null hypothesis. Observing the coefficient of the regression table, we can see Black or African American employees earn $5.79 less than white employees, and Hispanic or Latino earn $6.95 less than white employees. Asian employees, however, make $3.24 more than white employees, and Unspecified employees earn $8.39 more than white employees.

Evaluation

Evaluating Results

- We better understand Austin, Texas's workforce demographic when analyzing our data models. The Black or African American (28.26%) and Hispanic or Latino (13.05%) workforce make up 41.31% of the total workforce in Austin.
- However, 41.31% of the workforce does not look for or qualify for all positions. We further looked into this in our data model, exploring the staffing level within each ethnicity. Within the Hispanic or Latino workforce, 24.25% are workers, 2.23% are supervisors, and 1.41% are Managers. Within the Black or African American workforce, 10.65% are workers, 1.24% are supervisors, and 0.88% are Managers. This information is crucial to providing specific job opportunities.
- Both the Black or African American and Hispanic or Latino populations have a large portion of their workforce, between 30 to 59 years old. However, the supervisor and manager staffing levels between the age group of 30 to 39 have the opportunity for growth compared to the age group of 40 to 49, which have more than double supervisors and managers.
- The Black or African American and Hispanic or Latino workforce are underrepresented within the less populated positions within our data models. The data models have explicitly focused on IT and Engineering position titles, as many jobs in Dell are IT or Engineering focused.
- Learning the average hourly rate is crucial to providing career opportunities to the Black or African Americans and Hispanic or Latino. Providing a higher-paying position is an incentive for individuals trying to climb the social ladder, which is the whole point of this project. Black or African Americans and Hispanics or Latino are being paid less compared to their white counterpart. Around $5 less.

Determine Next Steps

- Based on our results, the following steps for Dell Technologies is to create IT and Engineering positions mainly targeting the age group of 30 to 39 with a higher average hourly rate equal to their white counterparts. These positions also include supervisor and manager titles.

Deployment

We can surmise that our focus should be on potential employees aged 30 to 39. It has one of the highest volumes of employees, but it is also an age range that has enough experience and youth.

The first course of action is to create more IT and engineering jobs that will appeal to this age range. As noted earlier, in general, regardless of ethnicity, there is a large volume of individuals between the ages of 30 through 39 and 40 through 49. It would be in the best interest of Dell to take advantage of this data and focus on using this data to achieve its current and future goals. We plan to incorporate this plan into fruition by collaborating with human resources and marketing departments to market these new positions with higher or equal pay to be more appealing to individuals of that age range—for example, college career fairs, Handshake, LinkedIn, Indeed, and other avenues.

It would prove fruitful if we increased or equal the pay amongst all baseline IT and engineering. The reasoning behind this is to address the issue of unequal pay between individuals who register as and who register as African American and Latino. This would prove to be more of an incentive for applicants of all age ranges to apply at Dell Technologies, let alone ages between 30 to 39. Regarding the issue of increasing the number of IT and engineering positions, it should not be an issue. Due to today's climate, both industries are not only constantly growing but are constantly sought after. Dell themselves are continually looking to hire more individuals in those positions anyway to compete with their competitors and further future goals.