Skills for the City:

**QUICK VIEW**

|  |
| --- |
| Company: |
| Role Played: Researcher |
| Learning Objectives: |
| Technical Skills   * + SQL   + R   + Data visualization |
| Soft Skills   * Presentation skills * Group work skills * Data contextualization skills   Brief Case Summary:The following xCase will require you to analyze labor market data and make appropriate recommendations on how workforce funding should be spent to remediate the difficulty employers have filling roles.  Desired Deliverables:   * Script used to create and query database * Any other relevant scripts or files * Strategy presentation that demonstrates the following:   + Methodology overview   + Results   + Recommendations |
|  |
|  |

Charting Data-Driven Strategies to Combat New York City’s Skills Gap

A LevelEdu xCase Developed in conjunction with Burning Glass Technologies

1. Background

1.1 What is an xCase?These engaging lab exercises will require you to apply the skills you are gaining as a part of your Level coursework to a research or business problem that is rooted in a real organization. After exploring and analyzing datasets provided by these companies, you will derive insights from the data and create a work product that will be presented to the rest of the class.

1.2 Your company: [Burning Glass](http://burning-glass.com), a leading labor market analytics firm, is playing a growing role in informing the global debate on education and the workforce. Founded in 1999 with the goal of developing the world’s leading technologies for matching people with jobs, Burning Glass’s technologies deliver insight across workforce and economic development, career exploration and counseling, and job matching. Furthermore, Burning Glass solutions drive critical workforce, education, and economic development initiatives for more than a dozen state and national governments, as well as numerous educational institutions, regional agencies, global recruitment agencies, major employers, and job boards.

1.3 Problem statement:

This xCase will introduce you to how various regions in the United States experience and respond to the national skills gap. Ultimately, you will be asked to hone in on one region in particular—New York City—and make data-driven suggestions as to how community leaders can effectively combat this problem locally.

The “skills gap” has been defined as the “perceived mismatch between the needs of employers for skilled talent and the skills possessed by the available workforce” (JPMorgan Chase & Co. & Economic Modeling Specialists International, n.d., p. 1). Despite unemployment rates in the United States, “businesses routinely say they can’t find the skilled workers they need” (JPMorgan Chase & Co. & Economic Modeling Specialists International, n.d., p. 1).

This “gap” manifests differently throughout the country, due to variation in where certain industries are active and certain employers are headquartered. Access to educational opportunities also makes a difference in the size of a region’s skills gap. Thus, communities throughout the country have developed different interventions to combat the skills gap based on the unique challenges facing their particular region.

Though each region faces its own individual skills gap, there are some general trends that have been noted nationwide, such as industries in which demand is growing and the baseline education needed to be competitive in the labor market. Paramount among these trends is the fact that nearly 60 percent of today’s jobs require some sort of postsecondary training, and demand for middle skills jobs in the healthcare and technology fields is “exploding” (Symonds, Schwartz & Ferguson, 2011, p. 3). These jobs often only require an associate’s degree. Because of these trends, interventions aimed at combating a region’s skills gap often revolve around finding cost- and time-efficient ways to ensure there is access to in-demand postsecondary training opportunities.

Teams tasked with developing these interventions typically include (but are not limited to): state labor and development departments, workforce investment boards, private industry councils, higher education consortiums, think tanks, and employers. Additionally, data scientists and analysts—such as those from your company, Burning Glass—are critical fixtures in these teams as they possess the expertise to quantify and define the scope of the issue through capturing and making sense of labor market and workforce datasets.

In 2014, Burning Glass, along with Economic Modeling Specialists International (EMSI), provided the data analytics support for a pivotal whitepaper regarding New York City’s skills gap entitled, “[Closing the Skills Gap: Preparing New Yorkers for High-Growth, High-Demand, Middle-Skill Jobs](https://www.jpmorganchase.com/corporate/Corporate-Responsibility/document/54841-JPMC-GAP-REP-AW6.pdf).” The report was a part of [JP Morgan Chase & Co.’s New Skills at Work campaign](https://www.jpmorganchase.com/corporate/Corporate-Responsibility/new-skills-at-work.htm), a $250 million corporate responsibility effort aimed at analyzing and developing thoughtful solutions to the nation’s skills gap.

Since the publication of the report, Burning Glass has been asked by the task force addressing New York City’s skills gap to remain active in the effort, aptly named the NYC Skills Coalition (NYCSC). The NYCSC is particularly interested in Burning Glass’s ability to capture and make sense of rich datasets related to the local workforce and labor market. Furthermore, they are impressed by the organization’s ability to suggest strategies that are rooted in both the findings from their data analysis and what is currently possible given various stakeholders’ resources and capacity.

The NYCSC has just learned that $100 million will be available to put toward workforce development efforts in New York City over the next ten years. These funds derive from a mixture of federal, state, and private foundation grants and these parties have requested a detailed roadmap as to how the funds will be spent. To complete this roadmap, the NYCSC has asked Burning Glass to evaluate the City’s skills gap in detail, answering the questions listed in Section 2.2.

2. Approach

To complete the request that your team has been tasked with, you should abide the following steps. You will to clean and manipulate the dataset with which you have been provided. Then, during the final days of the xCase, you will determine which interventions would be best to suggest to the team and why. A suggested approach is shown below, however feel free to incorporate other data or explore

2.1 Week 1: Examining and importing the data, creating the database

* Read the xCase from beginning to end to understand what data that may be of interest to answer the business question
* Examine the dataset – can you open it in Excel? R? What about a text editor?

***Data: https://www.dropbox.com/s/1yq6p59ng26ozjx/BGT.txt?dl=1***

* + Understand what the data in each field represents
  + Determine the entities, which attributes make sense with which entities
  + Draw a preliminary entity-relationship diagram that includes:
    - Entities
    - Attributes
    - Relationships/Cardinality
    - Primary and Foreign keys

**Database creation options**

Please choose an option to create your database. If you find yourself spending more than 25% of the allotted time for the project on a more challenging approach, please try the next step down. The four options are as follows:

* **Most challenging** – I want to figure this out on my own
* **Challenging** – I want to figure this out with some hints on data exploration and cleaning
* **Moderate to Challenging** – Give me the clean data and an ERD to match, I will build the database based on that
* **Moderate** – Give me the database and an ERD to match, I want to spend more time on data visualization and analysis than database design and implementation

**Most challenging – I want to figure this out on my own**

* Explore the data and perform any necessary data cleaning
* Import the data using MySQL workbench as a single table, below are questions to guide this process:
  + What data types should you attribute to each column? What should be the default values for these fields?
  + Create a table in MySQL Workbench with fields corresponding to the .txt file
  + Create a .txt file with 200 rows for testing.
  + Which column are you using to uniquely identify each row? Why?
  + Import the truncated dataset, how many records are there? (Check that it is 200 records)
* Use your SQL skills to create an efficient relational database from this table ***based on your ERD***, developed above.
  + How efficient is your proposed database? Is there a lot of repeated data in each column, and does your entity-relationship diagram address this?
  + Create a final database to store the data, this should include all:
    - Tables
    - Attributes
    - Relationships/Cardinality/Primary and Foreign keys
  + Run a few simple queries to make sure your database is behaving as expected, then repeat this process with the full dataset to create the full database

**Challenging – I want to figure this out with some hints on data exploration and cleaning**

* Explore the data and perform any necessary data cleaning, hints:
  + There are some quirks to this dataset:
    - A job ID can be associated with multiple skills, does this make sense, what do you think?
    - A job ID can be associated with multiple certifications, does this make sense, what do you think?
    - A job ID can be associated with multiple degrees and majors, does this make sense, what do you think?
    - A job ID that is associated with the exact same set of skills, certifications, degrees and majors can have multiple values for salary and posting duration. Do you think this refers to the same job or a different job? (Hint: there is no definite answer, you need to make an assumption here)
      * If you believe those are duplicate postings, you will need to remove them
      * If you believe those are distinct postings, you will not be able to use jobID as the primary key in a jobs table (Hint: composite primary key or auto increment a new key)
    - Make any necessary changes to your entity relationship diagram based on the above.
* Import the data using MySQL workbench as a single table, below are questions to guide this process:
  + What data types should you attribute to each column? What should be the default values for these fields?
  + Create a table in MySQL Workbench with fields corresponding to the .txt file
  + Create a .txt file with 200 rows for testing.
  + Which column are you using to uniquely identify each row? Why?
* \* Import the truncated dataset, how many records are there? (Check that it is 200 records or 2293266 when you import the full dataset)
* Run a few simple queries to make sure your large table is behaving as expected
* Consider how the N/A or missing values are stored in each table and consider updating them to NULL
* Use your SQL skills to create an efficient relational database from this table ***based on your ERD***, developed above.
  + How efficient is your proposed database? Is there a lot of repeated data in each column, and does your entity-relationship diagram address this?
  + Create a final database to store the data, this should include all:
    - Tables
    - Attributes
    - Relationships/Cardinality/Primary and Foreign keys
  + Run a few simple queries to make sure your database is behaving as expected, then repeat this process with the full dataset to create the full database (beginning again at \*)

**Moderate to Challenging – Give me the clean data and an ERD to match, I will build the database based on that**

* If you explored the data, you would find:
  + There are some quirks to this dataset:
    - A job ID can be associated with multiple skills, does this make sense, what do you think?
    - A job ID can be associated with multiple certifications, does this make sense, what do you think?
    - A job ID can be associated with multiple degrees and majors, does this make sense, what do you think?
    - A job ID that is associated with the exact same set of skills, certifications, degrees and majors can have multiple values for salary and posting duration. Do you think this refers to the same job or a different job? (Hint: there is no definite answer, you need to make an assumption here)
      * Assume you believe those are distinct postings, you will not be able to use jobID as the primary key in a jobs table (Hint: auto increment a new key, called ID, as shown in the entity relationship diagram)
* Various ID columns have been added to another version of the dataset, linked here:

***https://www.dropbox.com/s/kuev6x8s0s0hxfj/BGT\_class.csv?dl=1***

* Use this data to create a database based on the following entity-relationship diagram: ***https://www.dropbox.com/s/u8q4d43a0b8mhkv/BGT\_class\_ERD.png?dl=1***
* Import the data using MySQL workbench as a single table, below are questions to guide this process:
  + What data types should you attribute to each column? What should be the default values for these fields?
  + Create a table in MySQL Workbench with fields corresponding to the .txt file
  + Create a .txt file with 200 rows for testing.
  + Which column are you using to uniquely identify each row? Why?
* \* Import the truncated dataset, how many records are there? (Check that it is 200 records, or 2293266 when you import the full dataset)
* Run a few simple queries to make sure your large table is behaving as expected
* Consider how the N/A or missing values are stored in each table and consider updating them to NULL
* Use your SQL skills to create an efficient relational database from this table ***based on your ERD***, developed above.
  + How efficient is your proposed database? Is there a lot of repeated data in each column, and does your entity-relationship diagram address this?
  + Create a final database to store the data, this should include all:
    - Tables
    - Attributes
    - Relationships/Cardinality/Primary and Foreign keys
  + Run a few simple queries to make sure your database is behaving as expected, then repeat this process with the full dataset to create the full database (beginning again at \*

**Moderate – Give me the database and an ERD to match, I want to spend more time on data visualization and analysis than database design and implementation**

* If you explored the data, you would find:
  + There are some quirks to this dataset:
    - A job ID can be associated with multiple skills, does this make sense, what do you think?
    - A job ID can be associated with multiple certifications, does this make sense, what do you think?
    - A job ID can be associated with multiple degrees and majors, does this make sense, what do you think?
    - A job ID that is associated with the exact same set of skills, certifications, degrees and majors can have multiple values for salary and posting duration. Do you think this refers to the same job or a different job? (Hint: there is no definite answer, you need to make an assumption here)
      * Assume you believe those are distinct postings, you will not be able to use jobID as the primary key in a jobs table, an auto incremented ID column has been created in this database as the primary key for the jobs table.
* Various ID columns have been added to another version of the dataset and a database has been based on the following entity-relationship diagram:

***https://www.dropbox.com/s/u8q4d43a0b8mhkv/BGT\_class\_ERD.png?dl=1***

* There are two options for using the database:
  + Use a read-only version of the database, called BGT\_CLASS, already on the Level server. Please refresh your list of databases, then let the Level staff know if you cannot see this database.
    - Run a few simple queries to make sure your database is behaving as expected. You can only SELECT and CREATE VIEW on this read-only copy. Please only CREATE VIEWs that begin with your username, as other students are accessing the same database. Views without a username will be removed.
    - ***Note: This does not work on NU-Wave Guest!***
  + Import the database using the dump file below. Simply open the file as you open .sql scripts, MySQL say the file is large and may suggest that you should run it without opening it. Run it, making sure to name your database appropriately.

*https://www.dropbox.com/s/u6ce5zlwdl71l8s/DumpOfBGTclass2.sql?dl=1*

* + - Run a few simple queries to make sure your database is behaving as expected.

2.2 Week 2: Exploring the data and addressing the problem statement, considering only NYC

* You will query the database to answer the following questions, save these datasets for later visualization:
  + Which states are represented in the data? In what percentages?
  + What is the distribution of posting duration?
  + Is the posting duration distribution similar across states?
  + Is the posting duration distribution similar across occupations?
  + Perform any other additional data exploration and document what you learn.
* What do you think may indicate a "skills gap" or "unfilled demand"? Number of listings? Posting duration? Something about skills data? A combination? Answer the following questions, and plan how you will visualize these results:
  + What are the top listed occupation names as percent of total listings?
  + Which occupation names have the longest durations?
  + In which fields or industries is there the greatest unfilled demand (by your definition)?
  + Within these fields, for which positions is there the greatest unfilled demand (by your definition)?
  + What are the demographics of these positions—specifically:
    - What level of education is required to perform these jobs?
    - Which particular skills are required to perform these jobs?
    - How many years of experience are required to perform these jobs?
    - Are there any special certifications required?
* Optional: Consider these same questions in other regions, and draw comparisons.
* Optional: Anything else that you want to consider?

2.3 Weeks 3 and 4: Visualize results and make recommendations, create your presentation

* Complete any additional data analysis. (e.g. if you analyzed other regions: In what ways are the skills gaps present in each region alike or different?)
* What are your findings, insights and conclusions?
  + Finally, based on your findings, which of the following options do you think would be the best use of workforce development funds?

1. Development of career pathways beginning in high school that provide students exposure to and preparation for in-demand industries and jobs. *Explore Pathways to Prosperity and the early college model.*
2. Community college scholarships that allow more students access to training for middle-skills jobs. *Explore offerings at local community colleges.*
3. Professional development programming that would help employers “skill-up” the workforce they already have. *Compare the number of workers available at every level to the number needed.*

* Visualize the data and your results using R or Tableau. ***An important note when importing databases (and tables) into Tableau – try planning your visualizations in SQL by creating VIEWs. Aggregate the data appropriately in different views, then import each VIEW into a DIFFERENT Tableau workbook. This will reduce the likelihood of Tableau crashing.***

***Note: when you "JOIN" different database entities in Tableau (rather than using VIEWs), all dimensions and measures will need to be re-aggregated carefully for visualization. This is because Tableau is implementing all of the JOINs as you've specified in the data source tab, so you end up with the flat file, full of duplicates.***

* + Create clean, informative and relevant visualizations to support your funding allocation solution, including but not limited to:
    - Industry demand
    - Position demand
    - Demographics
    - Challenge: Any additional visualizations required to further persuade your audience of your proposal.
* **Create a 10-12-minute presentation that does covers the following:** 
  + Brief overview of methodology, show your entity-relationship diagram and describe your database (especially if you created your own)
  + Outline your findings and insights regarding the skills gap in the New York City MSA.
  + Present your strategy to allocate workforce development funds toward initiatives where they will have the greatest impact.

3. Your Toolkit

3.1 The Dataset:

Burning Glass maintains a detailed database of online employer demand, which includes nearly 100M million current and historical job postings. The organization collects postings from over 38,000 online job sites to develop a comprehensive, real-time portrait of labor market demand. They further supplement with other sources such as the Bureau of Labor Statistics, Census Bureau, etc.

Each listing in the database contains information on each job such as title, employer and industry as well as job titles, skills, and qualifications that employers are seeking. The dataset you will be working with contains information from postings for jobs throughout the Northeastern United States. It contains more than two million rows. You will be focused on the data that relates to the New York City metropolitan statistical area (MSA), but please feel free to utilize data from the other MSAs as comparison or to further contextualize/complicate your strategy suggestions.

3.2 Supplemental Research Materials:

* [Whitepaper: “New Skills at Work: Closing the Skills Gap”](https://www.jpmorganchase.com/corporate/Corporate-Responsibility/document/54841-JPMC-GAP-REP-AW6.pdf)
* [Whitepaper: “Pathways to Prosperity: Meeting the Challenge of Preparing Young Americans for the 21st Century”](https://dash.harvard.edu/bitstream/handle/1/4740480/Pathways_to_Prosperity_Feb2011-1.pdf?sequence=1)
* [Jobs for the Future](http://www.jff.org/)