OSHA Data Clean-Up – MA 415 Midterm

William Dean

Contents

Objective

The OSHA data presents Health & Safety Regulations and all workplace injuries and accidents. This data is ready and available, however it is very messy and is not able to be interpreted easily. The goal of this project is to clean this OSHA dataframes given with the intent of looking for the most dangerous place to work in the state of Massachusetts and provide insite of the data through Data Exploration.

Problems with the OSHA Data – OSHA Clean-Up

Accident Table

The Accident Table from OSHA is a record of the accidents and the situation behind each accident. There is a lot of information as can be seen in the display of it below but it needs to be cleaned up to before exploring it and anything more.

ACTIVITY	SITE	STATE	NAME	RELINSP	SEX	DEGREE	NATURE	BODYPART	AGENT	EVENT	ENVIRON	HUMAN	TASK	HAZSUB	OCC_CODE	OCC
10096592	MA	NA	10096592	NA	3	21	04	16	08	07	06	2	NA	000	0	
10096592	MA	NA	10096592	NA	3	21	04	16	08	07	06	2	NA	000	0	
10096592	MA	NA	10096592	NA	3	21	04	16	08	07	06	2	NA	000	0	

Unnecessary Atributes

At first glance, the dataframe that represents the accidents in MA is very messy and has many attributes that would not be helpful to looking into the most dangerous places to work in Massachusetts, so these were the first to be eliminated. A Person's name or the RELINSP (An inpections key) doesn't give much use to look at dangerous locations and are taken out to simplify the data.

ACTIVITY	SEX	DEGREE	NATURE	BODYPART	SOURCE	EVENT	ENVIRON	HUMAN	TASK	HAZSUB	OCC_CODE	AGE
10096592	NA	3	21	04	16	08	07	06	2	NA	000	0
10096592	NA	3	21	04	16	08	07	06	2	NA	000	0
10096592	NA	3	21	04	16	08	07	06	2	NA	000	0

Getting Rid of Incorrect Information (NA Values)

Although this is better, there is much information that is stored as placeholder values when they should be NA. For example, an OCC_CODE of "000" does not refer to an Occupation and should be NA. Taking out all instances of bad values cleans the data more.

ACTIVITY	SEX	DEGREE	NATURE	BODYPART	SOURCE	EVENT	ENVIRON	HUMAN	TASK	HAZSUB	OCC_CODE	AGE
10096592	NA	3	21	04	16	08	07	06	2	NA	NA	NA
10096592	NA	3	21	04	16	08	07	06	2	NA	NA	NA
10096592	NA	3	21	04	16	08	07	06	2	NA	NA	NA

Give Appropriate Labels

There is a lot of information in the table that is not stored in this table but in other tables that would be better suited to have all in one location. Bringing this information to one table makes it easier to understand what the data means and will make exploration and analysis easier. The resulting table is much cleaner and can be seen below.

ACTIVITY	SEX	DEGREE	TASK	AGE	OCCUPATION	HAZSUB	HUMFACT	BODYPART	NATURE	SOURCE	ENVIRON	EVENT
10096592	NA	Non-Hospitalized	Irregular	NA	NA	NA	EQUIP. INAPPROPR FOR OPERATION	BODYSYSTEM	OTHER	FIRE/SMOKE	GAS/VAPOR/MIST/FUME/SMOKE/	INHALATION
10096592	NA	Non-Hospitalized	Irregular	NA	NA	NA	EQUIP. INAPPROPR FOR OPERATION	BODYSYSTEM	OTHER	FIRE/SMOKE	GAS/VAPOR/MIST/FUME/SMOKE/	INHALATION
10096592	NA	Non-Hospitalized	Irregular	NA	NA	NA	EQUIP. INAPPROPR FOR OPERATION	BODYSYSTEM	OTHER	FIRE/SMOKE	GAS/VAPOR/MIST/FUME/SMOKE/	INHALATION

Final Version

This provides much a much cleaner version of the original accident table but there exists rows which have the same values everywhere that are redundant. Taking them out givens a final version of the table we started with.

ACTIVITY	SEX	DEGREE	TASK	AGE	OCCUPATION	HAZSUB	HUMFACT	BODYPART	NATURE	SOURCE	ENVIRON	EVENT
10096592	NA	Non-Hospitalized	Irregular	NA	NA	NA	EQUIP. INAPPROPR FOR OPERATION	BODYSYSTEM	OTHER	FIRE/SMOKE	GAS/VAPOR/MIST/FUME/SMOKE/	INHALATION

ACTIVITYNO	SEX	DEGREE	TASK	AGE	OCCUPATION	HAZSUB	HUMFACT	BODYPART	NATURE	SOURCE	ENVIRON	EVENT
305548745	M	Fatal	Regular	47	ELECTRICIANS	NA	INSUF/LACK PROT WRK CLTHG/EQUIP	[illegible]	OTHER	ELEC APPARAT/WIRING	OTHER	FALL(FROM ELEVATION)
306811928	M	Fatal	Regular	39	ELECTRICAL POWER INSTALLERS AND REPAIRERS	NA	OTHER	NECK	OTHER	OTHER	OVERHEAD MOVING/FALLING OBJ AC	CAUGHT IN OR BETWEEN

OSHA Table

The OSHA Table contains the information for all of the Inpections that have happened which is able to give us more information about the accidents in Massachusetts. Our goal is to combine the important information from the OSHA table to the Accident table we have already made to how a condensed table with all the accident information. However, we run into a lot of the same problems as the Accident table so the OSHA table has to be cleaned and preped before joining. The given OSHA table is too large and messy and can be seen below:

CONTFLAG	HISTFLAG	OSHA1MOD	SMOLARGE	PREVCITE	PREVACTIVITY	ACTIVITYNO	REPORTID	HO_JOBTITLE	OWNERTYPE	OPTREPORT	ESTABNAME	SITEADD
NA	H	19840221	NA	NA	0	10236776	0111100	NA	C	00000000	DUBE DRY WALL	RT 1 MAIN ST
NA	M	19910523	NA	NA	0	103393633	0111100	NA	I	00000000	KNOWLTON MACHINE CO.	NEW ENGLAND POWER, SALEM HARBO
NA	H	19880618	NA	NA	0	18750034	0111400	NA	NA	00000000	RENTAL & FROST	NA

Cleaning the Table

After taking only the information we want, getting rid of incorrect values which should be NAs, and relabeling the keys stored as integers with more discriptive and accurate labels, we come of with a version of the OSHA table that will provide important information about what happened at each inspection. A final version of the OSHA table before combining with the Accident table can be seen below and is easier to understand the information compared to how it was originally presented.

ACTIVITYNO	ESTABNAME	SITEADD	OWNERTYPE	OPENDATE	CLOSEDATE	NAICS	INDUSTRY	city	latitude	longitude	Decade
10236776	DUBE DRY WALL	RT 1 MAIN ST	NA	1983-12-15	NA	NA	PLASTERING, DRYWALL, AND INSULATION	NA	NA	NA	80s
103393633	KNOWLTON MACHINE CO.	NEW ENGLAND POWER, SALEM HARBO	Private	1990-07-17	1990-07-20	NA	INDUSTRIAL MACHINERY, NEC	Salem	42.51685	-70.8985	90s

ACTIVITYNO	ESTABNAME	SITEADD	OWNERTYPE	OPENDATE	CLOSEDATE	NAICS	INDUSTRY	city	latitude	longitude	Decade
18750034	RENTAL & FROST	NA	NA	1979-05-14	1979-05-14	NA	SHEET METAL WORK	NA	NA	NA	70s

Complete Accidents Table

We have two tables with information about the accidents in MA, but now we want to store it all in one table. Using a left join, we can keep all the information about the accidents from the accident table, but we can also incorporate any additional information that the OSHA table will give about each accident. Joining the tables results in one single table that has all the information from the data frames which would be helpful in exploring the most dangerous places in Massachusetts.

ACTSEXDEGREETASKAGEOCCUPATIONNATUREHUMANFACTORBODYPARTSOURCEENVIRONMENTESTABNAMESITEADDOWNERTYPEOPENDATECLOSEDATENAICSINDUSTRYcitylatitudelongitudeDecade
1009659 2 Non-Irregular Hospitalized NA NA NA EQUIP IN-AP-PROPR FOR OPERATION BODY SYSTEMS/MULTIPLE/VAPOR/FUME/SMOKE/DUST DYNAMIC CON-STRUCTION CO INC MAS-SACHUSETTS AVENUE 1980-05-02 1980-05-02 NONRESIDENTIAL CON-STRUC-TION, NEC 42.26 -71.16708 80s
305548745 M Fatal Regular 47 ELECTRICIANS MISUSE/MALF WRK CLTHG/EQUIP ELECTROCUTION APPARATUS (WIRING) L POWER SER-VICE INCORPORAT 6M TAUNTON STREET Private 2003-04-24 2003-04-28 ELECTRICAL CONTRACTORS ELECTRICAL WORK 71.32392 00s
306811928 M Fatal Regular 39 ELECTRICAL POWER IN-STALLERS AND REPAIRERS OTHER NECK OTHER MOV-ING/FALLING OBJ AC CAUGHT IN OR BETWEEN LINE CON-STRUC-TION, LLC POND ROAD Private 2004-05-07 2004-05-07 POWER AND COM-MUNI-CA-TION LINE AND RE-LATED STRUC-TURES CONSTRUCTION WATER, SEWER, AND UTIL-ITY LINES Plymouth 41.89468 70.61999 00s

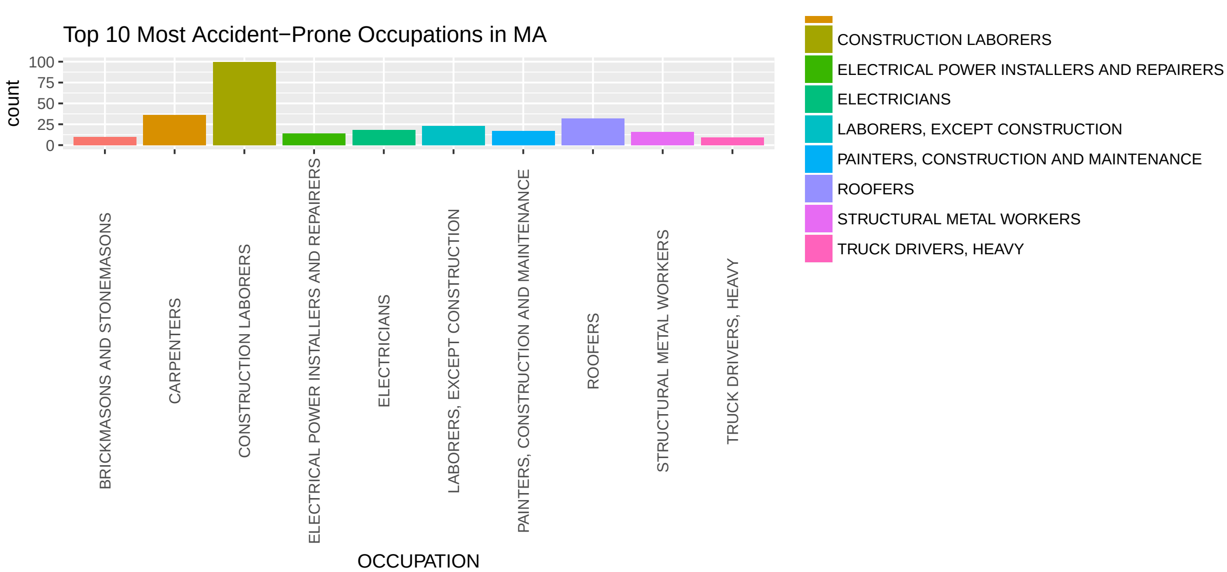
ACTIVITY	SEX	DEGREE	TASK	AGE	OCCUPATION	NATURE	HUMAN FACTOR	BODY PART	SOURCE	ENVIRONMENT	ESTAB NAME	SITE	OWNER	OPEN DATE	CLOSE DATE	SIC TEXT	INDUSTRY	city	latitude	longitude	Decade
306812744	M	Fatal	Regular	22	CARPENTERS	NA	POSITION INAPPROPRIATE FOR TASK	CHEST	HAND TOOL (POWERED)	FLYING OBJECT ACTION	[illegible] CONSTRUCTION, LLC	[illegible] CORWIN GROVE	Private	2004-05-18	2004-07-28	NEW MULTI-FAMILY HOUSING CONSTRUCTION (EXCEPT OPERATIVE BUILDERS)	RESIDENTIAL CONSTRUCTION, NEC	Plymouth	41.94468	-70.61999	2000s
305545279	M	Fatal	Regular	62	NA	NA	MISJUDGMENT, HAZ. SITUATION	[illegible]	[illegible]	MATERIAL HANDLING EQUIP./METHOD	[illegible] FIREWOOD	1350 WOODMAN'S WAY	Private	2002-11-19	2002-11-19	NA	WOOD PRODUCTS, NEC	Brewster	41.74665	-70.07773	2000s

Exploring the Data

Now that we have a Table with all the useful information about all the accidents that happened and have it organized, we can begin to explore the accidents in Massachusetts. Because of the our tidy data which brings in many types of attributes for each accidents, we can produce many visualizations from the data. Here are a few types:

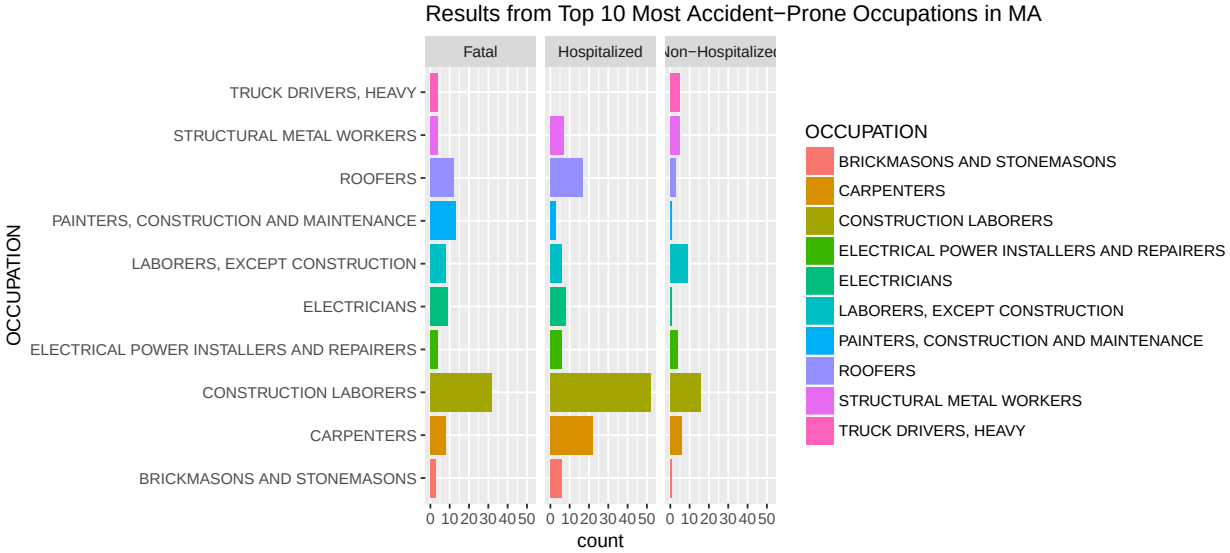
Bar Graphs

Since there are so many discrete variables, we can explore them with the use of bar graphs. Let's first consider the 10 most frequent **Occupations** in our data:



The plot above gives incite to which occupations lead to the most amount of injuries, but it does not tell us the severity of those injuries.

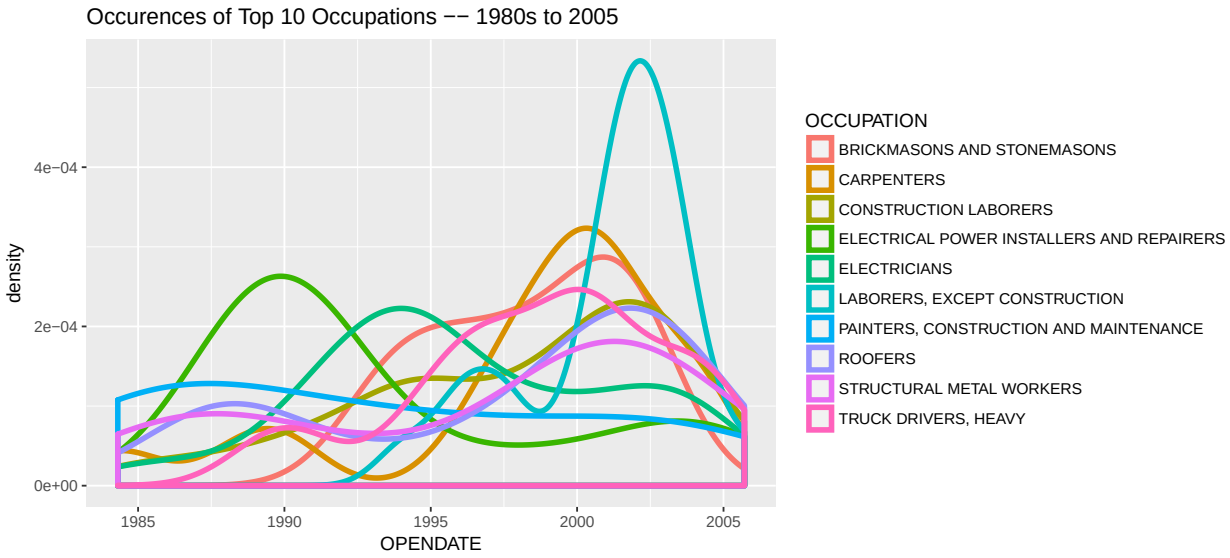
Because of that, consider the same set of data, but grouped by another variable, **Degree** of Accidents. Here we can consider the outcome of the 10 most frequent Occupations in the data:



We are able to consider if the most accident-prone jobs tend to have serious injuries or just many none serious injuries.

Time Series – Time Occurances

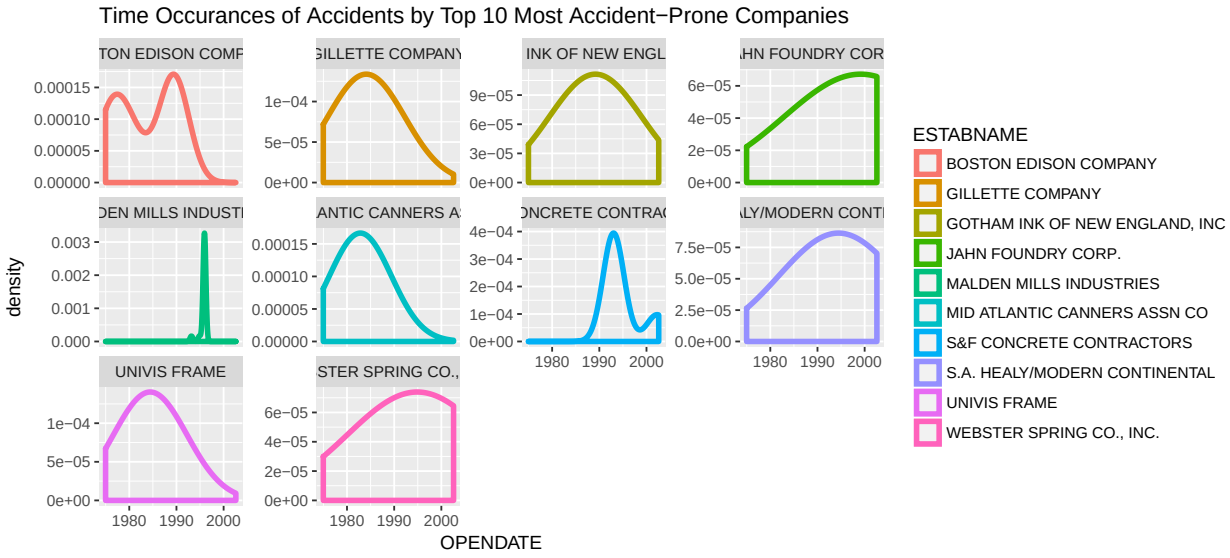
Using the Accident's Date, we can explore when and how many accidents happen. Consider still the Top 10 Accident-Prone Occupations. We can see when theses accidents happened:



Many of these results skew either left or right, so it may be considered if these are disappearing career paths or rising, more dangerous careers. Perhaps the job are changing.

Let's now explore another Attribute over time, Company names.

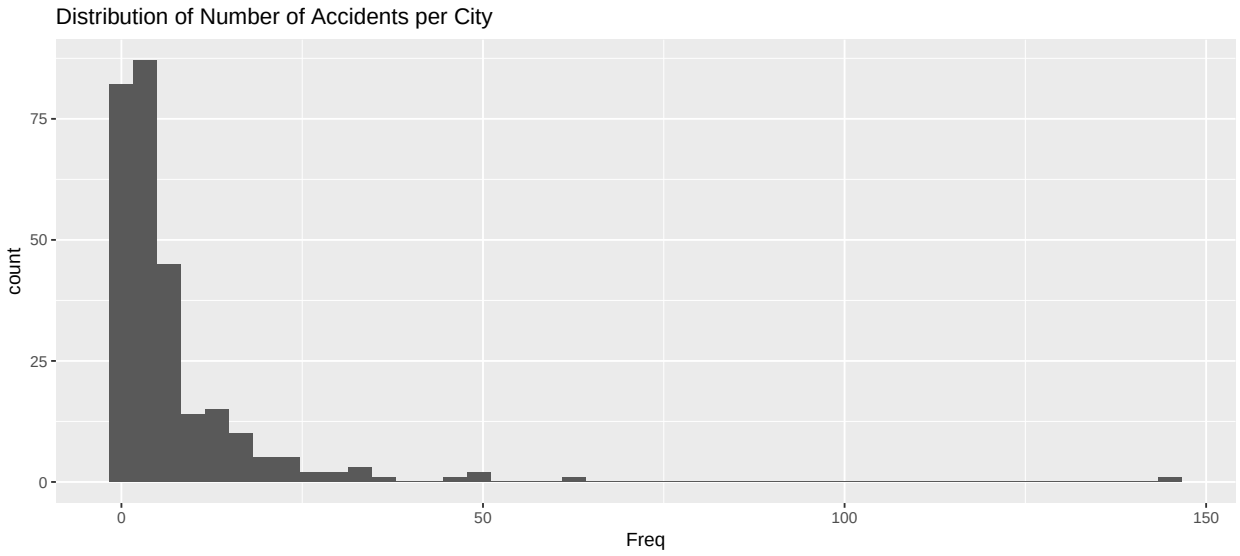
Sub-setting by the 10 Most Occuring Companies, we can see when in time these Accident-Prone companies had these Accidents:



All of these companies had many accidents but the time over which they had these accidents differ and can be visualized with the plots. Some have smaller accidents throughout time while some have large accidents that occur not so often.

Histograms

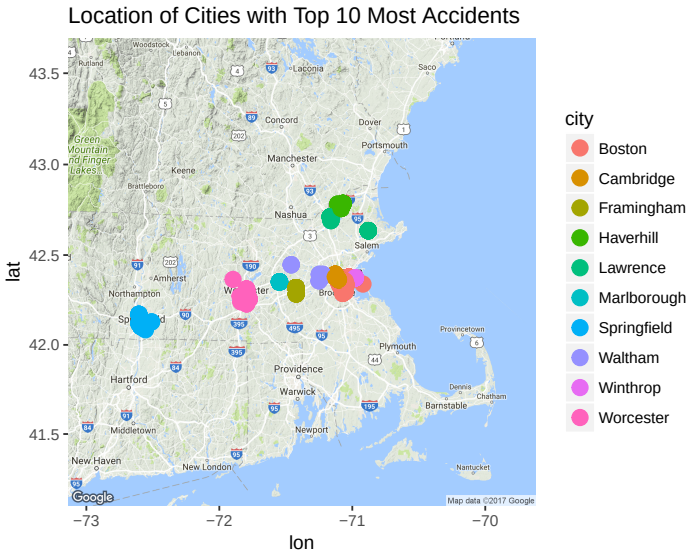
Histograms can be used to see the distributions of some of our discrete variables. Consider the accidents that occur per city in the whole time frame:



Over a time interval, we are able to see that lower amounts of accidents are more likely to occur. Cities with more accidents are less frequent and could be accounted by being larger, more industrial places.

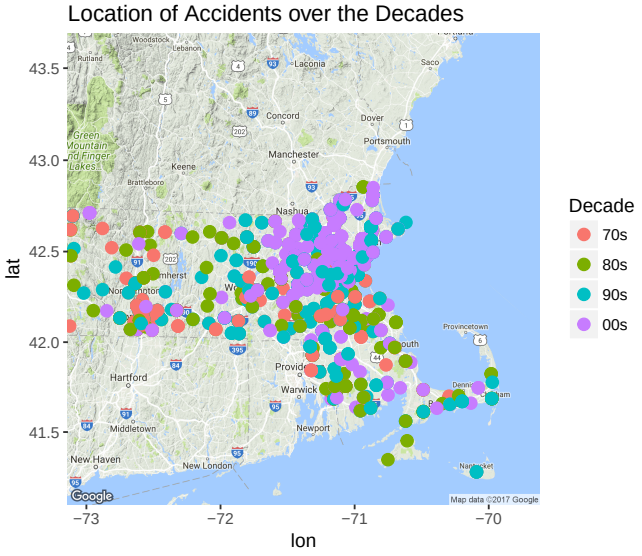
Map Graphics

With the use of location, we can visualize where in the state these accidents occured. We are able to see where the cities with the most accidents are.



We can see that many of these location are near Boston, but also near cities where there is a lot of industry compared to smaller places.

Another option is to see where in Massachusetts had the most accidents throughout time. To make it easier to see, let's group accidents by the decade in which they occured. We can see whether accidents are moving to new locations throughout the state.



There could be a trend that is moving accidents toward Boston.

Final Remarks

By cleaning up the data but also bringing in the most information we can about each accident, we get one data frame that is easy to understand and ready to explore. At the beginning, our data of interest was scattered across many tables and had many values that were incorrect but, by using R, we are able to bring the data to one place.

The data is versatile in the many ways it can be visually explored. The ways to explore the data that are above are just a subset of all possible. The data can be see through many of the dicrete variables, the frequency of them, the time occurance of accidents, and/or geographical locations of accidents.

The data that was prepared can be taken in many directions to discover the most dangerous place to work, and to uncover the multifaceted stories behind each.