Feature Engineering with Storm Events

Table of Contents

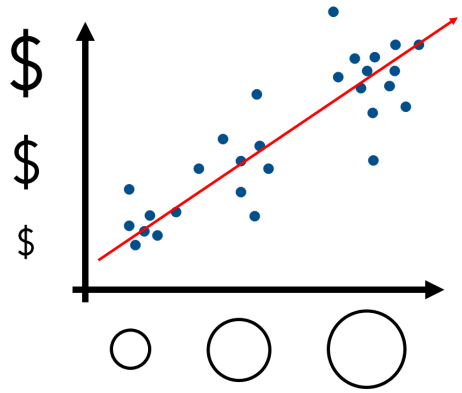
Introduction

In this reading you will import storm event data, extract the text for event descriptions along with the corresponding damage cost values, pre-process the data, find and extract features in the text to use for predicting damage, evaluate the features visually as well as ANOVA testing, and subsequently refine them further.

Of the many factors that affect how much damage a hail storm causes, hail size is probably the best predictor. However, the storm event data doesn't include hail size as a variable. Rather, there are text descriptions of the storm events containing qualitative descriptions of hail size. Here you can see several examples that use common items for size reference instead of giving numerical measurements.

> "Golf ball size hail broke rain gauge.
> Estimated time of report from radar."
>
> "Quarter size hail was reported in Daytona
> North by an off duty Fire Rescue Captain."
>
> "Penny size hail was reported 4 miles
> southwest of Greenbelt Lake."

Item names with known sizes in the descriptions of hail storms are features that you could use to predict damage.



Before you can start processing the text to identify useful features, you'll need to import it.

Note: as you proceed through this reading, it will be best to run each section individually in sequence.

Import data

You will need the storm events data from Exploratory Data Analysis (Course 1). The data could be imported using the Import Tool with some processing, but this time you can just use the code below. If you completed Course 1, you should be able to use your own import function as well. If you haven't already, make sure the folder containing the Storm Events files is on the search path.

```
events = importStormData("StormEvents_2016.csv")
```

```
events = 56003×16 table
```

...

	Month	Event_Type	Begin_Date_Time	Timezone	Injuries_Direct
1	July	Heavy Rain	2016-07-15 17:15:00	EST-5	0
2	July	Thunderstorm Wind	2016-07-15 17:25:00	EST-5	0
3	July	Thunderstorm Wind	2016-07-16 12:46:00	EST-5	0
4	July	Thunderstorm Wind	2016-07-08 17:55:00	EST-5	0
5	July	Thunderstorm Wind	2016-07-08 18:10:00	EST-5	0
6	July	Thunderstorm Wind	2016-07-08 19:10:00	CST-6	0
7	December	Winter Storm	2016-12-04 03:00:00	MST-7	0
8	December	Winter Storm	2016-12-04 04:00:00	MST-7	0
9	December	Winter Storm	2016-12-04 04:00:00	MST-7	0
10	July	Marine Thunderstorm Wind	2016-07-12 01:30:00	EST-5	0
11	July	Marine Thunderstorm Wind	2016-07-17 15:40:00	EST-5	0
12	August	Thunderstorm Wind	2016-08-16 19:23:00	EST-5	0

	Month	Event_Type	Begin_Date_Time	Timezone	Injuries_Direct
13	August	Thunderstorm Wind	2016-08-16 20:18:00	EST-5	0
14	August	Thunderstorm Wind	2016-08-13 20:15:00	EST-5	0
15	August	Heat	2016-08-13 10:00:00	EST-5	0
16	August	Excessive Heat	2016-08-13 10:00:00	EST-5	0
17	August	Thunderstorm Wind	2016-08-13 19:15:00	EST-5	0
18	August	Thunderstorm Wind	2016-08-13 19:45:00	EST-5	0
19	August	Thunderstorm Wind	2016-08-13 19:58:00	EST-5	0
20	August	Flash Flood	2016-08-02 03:00:00	EST-5	0
21	August	Thunderstorm Wind	2016-08-12 16:13:00	EST-5	0
22	August	Thunderstorm Wind	2016-08-13 18:39:00	EST-5	0
23	August	Thunderstorm Wind	2016-08-13 19:55:00	EST-5	0
24	August	Heat	2016-08-14 11:00:00	EST-5	0
25	August	Thunderstorm Wind	2016-08-13 19:50:00	EST-5	0
26	August	Thunderstorm Wind	2016-08-13 18:50:00	EST-5	0
27	August	Thunderstorm Wind	2016-08-13 19:15:00	EST-5	0
28	August	Thunderstorm Wind	2016-08-13 20:05:00	EST-5	0
29	August	Flash Flood	2016-08-02 03:15:00	EST-5	0
30	August	Flash Flood	2016-08-02 03:00:00	EST-5	0
31	August	Flash Flood	2016-08-02 03:15:00	EST-5	0
32	August	Heat	2016-08-12 10:00:00	EST-5	0
33	August	Thunderstorm Wind	2016-08-12 16:06:00	EST-5	0
34	August	Thunderstorm Wind	2016-08-13 20:00:00	EST-5	0
35	August	Heat	2016-08-14 11:00:00	EST-5	0
36	August	Heat	2016-08-14 11:00:00	EST-5	0
37	August	Heat	2016-08-14 11:00:00	EST-5	0
38	August	Thunderstorm Wind	2016-08-16 18:58:00	EST-5	0
39	March	High Wind	2016-03-16 12:00:00	CST-6	0
40	March	High Wind	2016-03-16 12:00:00	CST-6	0
41	February	Flood	2016-02-25 11:57:00	EST-5	0
42	February	Flood	2016-02-24 20:16:00	EST-5	0
43	February	Flash Flood	2016-02-24 18:30:00	EST-5	0
44	February	Flood	2016-02-24 23:36:00	EST-5	0
45	February	Flood	2016-02-24 21:37:00	EST-5	0

	Month	Event_Type	Begin_Date_Time	Timezone	Injuries_Direct
46	March	High Wind	2016-03-16 12:00:00	CST-6	0
47	March	High Wind	2016-03-16 13:00:00	CST-6	0
48	March	Marine High Wind	2016-03-16 13:00:00	CST-6	0
49	December	Drought	2016-12-01 00:00:00	EST-5	0
50	December	Drought	2016-12-01 00:00:00	EST-5	0
51	December	Drought	2016-12-01 00:00:00	EST-5	0
52	December	Drought	2016-12-01 00:00:00	EST-5	0
53	December	Drought	2016-12-01 00:00:00	EST-5	0
54	December	Drought	2016-12-01 00:00:00	EST-5	0
55	December	Drought	2016-12-01 00:00:00	EST-5	0
56	December	Drought	2016-12-01 00:00:00	EST-5	0
57	December	Drought	2016-12-01 00:00:00	EST-5	0
58	December	Drought	2016-12-01 00:00:00	EST-5	0
59	December	Drought	2016-12-01 00:00:00	EST-5	0
60	December	Drought	2016-12-01 00:00:00	EST-5	0
61	December	Strong Wind	2016-12-27 05:00:00	EST-5	0
62	December	Strong Wind	2016-12-27 02:00:00	EST-5	0
63	December	Strong Wind	2016-12-27 02:00:00	EST-5	0
64	December	Strong Wind	2016-12-27 03:00:00	EST-5	0
65	December	High Wind	2016-12-15 22:00:00	EST-5	0
66	January	Hail	2016-01-09 01:03:00	CST-6	0
67	December	Winter Weather	2016-12-17 00:00:00	EST-5	0
68	December	Winter Weather	2016-12-17 00:00:00	EST-5	0
69	December	Winter Weather	2016-12-17 00:00:00	EST-5	0
70	December	Strong Wind	2016-12-18 14:00:00	EST-5	0
71	December	Strong Wind	2016-12-27 02:00:00	EST-5	0
72	December	Strong Wind	2016-12-27 03:00:00	EST-5	0
73	January	Hail	2016-01-21 15:20:00	CST-6	0
74	January	Thunderstorm Wind	2016-01-21 15:22:00	CST-6	0
75	January	Hail	2016-01-21 15:35:00	CST-6	0
76	January	Thunderstorm Wind	2016-01-21 15:35:00	CST-6	0
77	January	Hail	2016-01-21 15:55:00	CST-6	0
78	December	Strong Wind	2016-12-15 12:00:00	EST-5	0

	Month	Event_Type	Begin_Date_Time	Timezone	Injuries_Direct
79	December	Strong Wind	2016-12-15 12:00:00	EST-5	0
80	December	Drought	2016-12-01 00:00:00	EST-5	0
81	December	Strong Wind	2016-12-27 03:00:00	EST-5	0
82	December	Strong Wind	2016-12-27 04:00:00	EST-5	0
83	January	Hail	2016-01-21 16:05:00	CST-6	0
84	January	Thunderstorm Wind	2016-01-21 16:07:00	CST-6	0
85	January	Hail	2016-01-21 16:10:00	CST-6	0
86	January	Thunderstorm Wind	2016-01-21 16:13:00	CST-6	0
87	January	Thunderstorm Wind	2016-01-21 16:37:00	CST-6	0
88	January	Hail	2016-01-21 17:31:00	CST-6	0
89	January	Thunderstorm Wind	2016-01-21 17:40:00	CST-6	0
90	January	Hail	2016-01-21 18:15:00	CST-6	0
91	January	Hail	2016-01-21 08:11:00	CST-6	0
92	January	Thunderstorm Wind	2016-01-21 14:23:00	CST-6	0
93	January	Hail	2016-01-21 14:30:00	CST-6	0
94	January	Thunderstorm Wind	2016-01-21 15:04:00	CST-6	0
95	January	Thunderstorm Wind	2016-01-21 18:35:00	CST-6	0
96	January	Hail	2016-01-21 18:40:00	CST-6	0
97	January	Thunderstorm Wind	2016-01-21 19:26:00	CST-6	0
98	January	Thunderstorm Wind	2016-01-21 19:23:00	CST-6	0
99	January	Thunderstorm Wind	2016-01-21 19:42:00	CST-6	0
100	January	Hail	2016-01-21 16:02:00	CST-6	0

⋮

The first time you go through this reading, you should just use *StormEvents_2016.csv*. However, if later you want to import all of the storm event files at once, you can use a datastore. In this case you'll need to do a few things:

1. Locate the file path on your system with the storm events data for Exploratory Data Analysis (Course 1). For example, "C:\Courses\Exploratory Data Analysis\StormEvents".
2. Remove *StormEvents_2017_finalProject.csv* from the directory since you won't want to duplicate data in *StormEvents_2017.csv*
3. Uncomment the code below and modify the file path "S:\Coursera Development\Exploratory Data Analysis\Course Files\StormEvents" to match the one you found in Step 1.

```
% ds = fileDatastore("C:\Coursera Development\Exploratory Data Analysis\ ...
%   Course Files\StormEvents", "ReadFcn", @importStormData, "UniformRead",true);
```

```
% events = readall(ds);
```

Now combine property and crop cost into a single total.

```
% Replace all NaN values with 0
events.Property_Cost = fillmissing(events.Property_Cost,"constant",0);
events.Crop_Cost = fillmissing(events.Crop_Cost,"constant",0);

events.Cost = events.Property_Cost + events.Crop_Cost;
```

Now extract hail events with non-missing entries using logical indexing and the & (AND) operator. Note that an empty string in the Event Narrative does not get flagged by `ismissing`, so you could pass `""` as an indicator, or use a function like `strlength` to ignore empty strings.

```
% Filter out the hail events
rows2keep = events.Event_Type=="Hail" & ~ismissing(events.Cost) & ...
    strlength(events.Event_Narrative) > 0;
cols2keep = ["Event_Narrative","Cost"];
events = events(rows2keep,cols2keep);


% Check how hyphens are used in event narratives
hasHyphen = contains(events.Event_Narrative,"-");
disp(events.Event_Narrative(hasHyphen))
```

```
    "Quarter size hail fell near I-220 and Highland Colony Parkway."
    "Meteorologist from the 26th Operational Weather Squadron at Barksdale Air Force Base reported half-dollar size
    "A trained weather spotter observed penny-sized hail falling near State Roads 50 and 429 in Ocoee."
    "Quarter size hail was reported 6 miles west-northwest of Hedley."
    "Quarter size hail fell 3 miles south-southeast of Lake McClellan."
    "Public reported quarter size hail near the intersection of Highway 64 and I-40."
    "Spotter reported quarter size hail off I-85."
    "Dime size hail fell near mile marker 173 on I-75."
    "Hail fell one mile east of the Silverton exit on I-77."
    "A report was received via social media of half-dollar size hail covering the ground in spots near Haines City."
    "A picture of an oblong, spiky hailstone was posted to the WALB-TV facebook page.  It was estimated to be about
    "Quarter to half dollar size hail fell in a swath from 2 miles south-southwest of Hillsborough to 2 miles east-s
    "Quarter size hail was reported 2 miles west-northwest of Henderson, near the intersection of Enon Road and High
    "Golf ball size hail was reported near Paschall, approximately 4 miles north-northeast of Wise."
    "Quarter to ping pong ball sized hail was reported along a swath from 5 miles south-southeast of Hope Mills to C
    "Nickel size hail fell in the Fairview Alpha Community along the Red River/Natchitoches Parish line. Pictures we
    "Quarter size hail fell in Winnfield. A photo of the hail was posted to the KALB-TV Facebook page."
    "Quarter size hail fell in Jena. A photo was posted to the KALB-TV Facebook page."
    "Broadcast media reported nickel-sized hail near the intersection of Hwy 84 and FM 1996."
    "A trained spotter reported quarter to half-dollar sized hail in Ireland, TX."
    "The hail ranged from nickel to quarter-sized."
    "A delayed report was received from KSN Channel 3 from a viewer showing a 3-inch diameter hailstone."
    "Hail ranging from quarter to golf ball-sized was accumulating."
    "Hail to the size of quarters was reported along a swath from 5 miles east-southeast of Hillsborough northeast t
    "Hail up to the size of golf balls was reported along a swath from Duke Homestead Boulevard near Highway 157 to
    "Ping-pong ball size hail fell 3 miles west of Waskom."
    "Hail up to the size of quarters was reported 2 miles south-southwest of Butner."
    "Hail up to half-dollar size was reported along a swath from 2 miles west-southwest of Durham to 2 miles north-n
    "Nickel to quarter size hail fell for 4-5 minutes in Hall Summit."
    "WJHG-TV relayed a report of quarter size hail near the intersection of Highways 20 and 77."
    "An Off-Duty NWS Employee reported that ping pong ball size hail fell at the Grand Bayou Resort at Grand Bayou R
    "Quarter-size hail reportedly covered the ground."
```

"A photo posted to the KTBS-TV Facebook page showed measured hailstones ranging from golfball size to just small
"Quarter size hail fell in the Bosco community in Southern Ouachita Parish near the Caldwell Parish line. Report
"The public posted a photo of quarter size hail that fell in Sterlington on the KNOE-TV Facebook page."
"Quarter size hail was reported on Cooktown Road one-half mile north of Interstate 20."
"Quarter to half dollar size hail fell in the Cartright community in Northeast Jackson Parish. Report from a pic
"A CO-OP observer reported dime to penny sized hail and wind gusts up to 45 mph."
"An Off-Duty NWS Employee reported that the hail grew to the size of golf balls at the Grand Bayou Resort at the
"Dime-size hail was reported in the Lake of Egypt area."
"Dime to penny size hail fell east of Heflin per KTAL-TV."
"Hail to the size of quarters was reported near the intersection of Meyers-Cemetery Road and W John Paul Jones F
"Amateur radio reported one-inch diameter hail approximately 6 miles north of Weatherford, TX."
"Amateur radio reported one-inch diameter hail approximately 6 miles north of Weatherford, TX."
"A photo was posted to the KSLA-TV Facebook page of hail slightly larger than golfballs that fell in Winthrop."
"Quarter-size hail was reported near the intersection of U.S. Highway 51 and Illinois 154 by a sheriff deputy ar
"Nickel to quarter-size hail fell in the Belle Prairie City area."
"Quarter size hail and winds to around 60 mph estimated on Interstate-35."
"Aside from the ping pong ball size hail, several 3-4 branches were snapped off of trees."
"A swath of dime to quarter sized hail hail occurred across the northern portion of the county. The largest hai
"Reported by the public to Kwch-Tv."
"The hail occurred on K-14 and was a delayed report received via Twitter. The time of the event was adjusted per
"Winds also estimated to be 30-40 mph."
"Half-dollar size hail reported by Emergency Manager."
"The hail occurred near the Kellogg/135th W Intersection and ranged from quarter to half dollar-sized."
"The off-duty NWS employee measured the hail near the NW 21st/W 119th Intersection. No damage was reported."
"The hail ranged from quarter to golf ball-sized."
"Public report of 1.00 inch hail from thunderstorm 5 miles west-southwest of Oildale California in Kern County w
"Dime sized hail fell on Avon-Darlove Road."
"Largest stones were three-quarters of an inch in diameter."
"The hail occurred at the I-135/Pawnee Interchange. No damage was reported."
"Hail up to the size of quarters was reported approximately 2 miles east-northeast of Summerfield."
"Several reports of dime to nickel sized hail were received from trained spotters and via the broadcast media wi
"Quarter to half dollar-sized hail was reported at the fire station just east of Conway Springs."
"Broadcast media relayed a report of 2-inch hail near Kress. Damage was not known."
"A swath of hail occurred across central Lincoln County. Golf ball sized hail fell at the intersection of Highwa
"The hail occurred near the I-235/S. Meridian Intersection, but no damage was reported."
"Penny to quarter sized hail occurred in a swath from near WDAM-TV studios in Eastabuchie to Moselle."
"Hail was estimated to be half dollar size at the KFDA studio, 6 miles north-northwest of Amarillo."
"A trained spotter reported quarter size hail and estimated 60 mph winds at Exit 96 on I-65 in Goodlettsville."
"Thunderstorms associated with a strong cold front moving across the area during the early morning hours on the
"Thunderstorms associated with a strong cold front moving across the area during the early overnight hours on th
"Pea to nickel size hail with a few golf ball size hailstones was reported one mile west-northwest of Stinnett."
"A low-precipitation supercell produced 3.5 inch diameter hailstones measured by storm chasers and found along H
"Hail up to the size of golf balls fell along a swath from 4 miles west of New Hope to 4 miles west-southwest of
"Hail up the the size of golf balls fell near Winston-Salem State University."
"The hail size during this period ranged from quarter to ping-pong ball in size."
"Amateur radio reported one-inch diameter hail in Weatherford, TX."
"Amateur radio reported ping-pong ball sized hail in the town of Weatherford, TX."
"Public measured one-inch diameter hail in the Crown Point area."
"Half dollar size hail occurred 11 miles east-northeast of Channing."
"Thunderstorms associated with an offshore low pressure system moved through Caroline County during the late aft
"Thunderstorms associated with an offshore low pressure system moved through Queen Anne's County during the late
"A truck driver on I-75 at exit 288 reported hail as big as golf balls."
"Half dollar size hail fell 7 miles west-southwest of Kellerville. The hail accumulated up to an inch deep."
"Thunderstorms associated with an offshore low pressure system moved through Queen Anne's County during the late
"Thunderstorms associated with an offshore low pressure system moved through Queen Anne's County during the late
"Thunderstorms associated with a cold front moving from west to east across the area during the late evening hou
"A public report indicated nickel sized hail 5 miles south-southeast of Weatherford, TX."
"A social media report indicated nickel-sized hail near Texas Motor Speedway."
"A social media report indicated quarter-sized hail near Trophy Club, TX."
"An amateur radio operator reported quarter size hail at State Highway 20 and I-75 northeast of Cartersville and
"Three-quarter inch diameter hail was reported in the Denton, TX surface weather observation."
"A public report estimated 2-inch diameter hail near the intersection of State Highway 16 and County Road 262."
"A public report estimated 2-inch diameter hail near the intersection of State Highway 16 and County Road 262."
"A swath of very large hail fell across central Muhlenberg County. Baseball-size hail was reported about a mile

"A social media report indicated hail slightly larger than quarter-sized just north of Weatherford, TX."
"Amateur radio reported quarter-sized hail approximately 2 miles northwest of Cresson, TX."
"A trained spotter reported penny size hail at I-75 and Highway 92."
"A public report estimated quarter-sized hail approximately 5 miles south of Santo, TX."
"Pea to penny size hail fell 3 miles west-southwest of Simms."
"Hail observed just north of the Sullivan-Knox county line."
"Quarter size hail was reported 1 mile north-northwest of Shamrock, with no accumulation."
"Quarter sized hail was reported at the intersection of FM 455 and TX 5-N in Anna."
"A storm chaser reported penny-sized hail on the south side of Grapevine Reservoir."
"Dime to nickel-size hail was reported in Herrin and just north of town."
"A television news crew videotaped quarter-size hail."
"Dime-size hail was reported in and just southeast of Evansville. Small branches less than an inch in diameter w
"Ping pong ball size hail occurred 1 mile east-southeast of Simms."
"Baseball size hail was reported 10 miles west-southwest of Wolf Creek Park."
"Half dollar size hail fell 3 miles west-northwest of Higgins."
"A slow-moving severe thunderstorm moved into western Wayne County, dumping quarter to ping-pong ball size hail
"A swath of large hail up to baseball sized hail occurred across Lamar County. This occurred as a supercell thun
"Young County Sheriff's Department reported quarter-sized hail in Graham, TX."
"A slow-moving severe thunderstorm crossed the city of Fairfield, producing a swath of quarter to golf-ball size
"Golf-ball size hail was reported along the White County line, about a mile east of Grayville."
"These ping-pong ball size hailstones were photographed."
"Trained spotter reported hail up to quarter-sized mixed in with smaller hail."
"Penny size hail was reported 3 miles south-southeast of Greenbelt Lake."
"Quarter-size hail occurred in a swath from Mount Vernon, east along Illinois Route 15, to the Wayne County line
"A trained spotter reported quarter size hail at the intersection of I-20 and State Highway 70."
"A Facebook report indicated quarter size hail and strong winds occurred about 4 miles south-southeast of Westmo
"Thunderstorms associated with a strong cold front moving across the area during the late evening hours on the 2
"The news media also reported quarter-sized hail on I-135 at mile marker 94."
"A major hailstorm continued east-northeast across the Owensboro area from western Daviess County. The largest h
"A severe thunderstorm produced a wide swath of very large hail across western Daviess County. Hailstones the si
"The public sent, via Twitter, a picture of quarter to half dollar-sized hail."
"The public also sent, via Twitter, a picture of the tennis ball-sized hail. Overall, hail had been occurring at
"The trained spotter reported the hail ranged from dime to quarter-sized."
"Numerous reports were received from trained spotters, broadcast media, and social media of large hail falling o
"The ping pong ball-sized hail covered the ground."
"The ping pong ball-sized hail covered the ground."
"No damage was reported. The Saline County Sheriff also reported quarter-sized hail at 12th and 9th Streets in S
"There was nickel-sized hail reported at Caspian."
"Several inches of hail accumulated on Highway 550 near Counselor. Slick travel conditions resulted in a head-on
"A slow moving storm produced a 12-mile stretch of severe hail through Lubbock ranging in size from quarters to
"Law enforcement as well as many others reported nickel to quarter-sized hail."
"EM reported up to 2-inch diameter hail caused considerable damage in the Sugar Hill area. At least one home rec
"Spotter reported golf ball to 2-inch diameter hail."
"A thunderstorm produced two and a half inch hail near I-10 and De Zavala Rd. in northwestern San Antonio."
"A public report indicated half-dollar sized hail in South McKinney."
"A trained spotter reported tennis ball sized hail 2 miles north-northeast of Plano, TX."
"A trained spotter reported golf-ball sized hail in the town of Lucas, TX."
"A trained spotter reported golf-ball sized hail near the intersection of Locust Dr and University St in Denton,
"A trained spotter reported golf-ball sized hail near the intersection of Parvin St and McCormick St in Denton,
"The public reported, via Twitter, that quarter-sized hail occurred in town. The report was relayed by KAKE Char
"A thunderstorm produced 2.75 inch hail near USAA at I-10 and Huebner Rd."
"Golf ball size hail fell 5 miles east-northeast of Felt."
"A trained spotter reported quarter-sized hail approximately 5 miles east of Denton, TX."
"A trained spotter reported quarter-sized hail in Corinth, TX."
"A trained spotter reported penny-sized hail about a mile north of Bridgeport, TX."
"A public report indicated golf-ball sized hail in Plano just northwest of Independence and Mcdermott."
"Nickel-size hail fell on Highway 70 a couple miles west of the Green River."
"Quarter-size hail fell in the same location that nickel-size hail fell earlier."
"Media reported 3/4 hail on Enola Rd near I-40."
"A trained spotter reported golf-ball sized hail in Myra, TX."
"A spotter in West Ishpeming reported dime-sized hail."
"A storm chaser for KFDA-TV in Amarillo reported quarter size hail southwest of Childress along US Highway 62."
"Fire and rescue reported golf-ball sized hail approximately one mile north of Alvord, TX."
"Amateur radio reported quarter-sized hail approximately 4 miles south of Slidell, TX."

"A NWS employee reported nickel-sized hail in Negaunee with a few hailstones just under an inch in diameter."
"A NWS employee observed dime-sized hail in Negaunee."
"There were public reports of dime-sized hail on Business M-28 between Ishpeming and Negaunee."
"There were public reports of dime-sized hail observed in Hardwood."
"There were reports via social media of estimated penny-sized hail or larger at the Negaunee North ball field."
"Penny-sized hail was observed in Gladstone for about 5 minutes."
"The Bark River Fire Department observed penny to nickel-sized hail."
"There was a delayed report via social media of penny-sized hail near Hogback Mountain. The hail was accompanie
"A severe thunderstorm produced half-dollar size hail in Bard."
"A trained spotter reported golf-ball sized hail near Highway 205 and Ralph Hall Parkway."
"Amateur radio reported two-inch diameter hail in Royce City, TX."
"A report was received via amateur radio indicating quarter-sized hail approximately 4 miles south of Sulphur Sp
"A trained spotter reported quarter-sized hail in Como, TX."
"Some hail stones from ping-pong ball to baseball size fell with the storm. Gusts to 40 mph also accompanied the
"A trained spotter reported nickel-sized hail near the intersection of Interstate 30 and Bobtown Road in Garland
"A trained spotter reported quarter-sized hail 8 to 10 miles north of Paris, TX."
"A public report indicated 2-inch diameter hail on Murphy Road in Parker, TX and also on Ranchview Ct."
"A trained spotter reported 2-inch diameter hail in the city of Plano, TX. Spotters reported 2 hail at Legacy Dr
"One-half to one inch diameter hail fell between downtown Grand Forks and the University of North Dakota."
"Occurred at I-470 and Gage Blvd. Wind gust was also estimated to be 40 mph."
"A public report indicated hen-egg sized hail south-southwest of the city of Emory, TX."
"Emergency management reported golf-ball sized hail near the Lake Tawakoni Dam near the county line."
"Hail size ranged from pea size on the south end of town to ping pong ball size on the north end. Highway 39 had
"Half dollar size hail was measured 3 miles west-southwest of Dalhart."
"Quarter size hail fell one mile south-southwest of Cactus."
"Spotters reported quarter to golf ball sized hail approximately one mile east-northeast of Copperas Cove. Damag
"A trained spotter reported one-inch diameter hail approximately 2 miles north of Sunset, TX."
"A trained spotter reported nickel-sized hail near Powderly, TX."
"The local post office reported golf-ball sized hail several miles north of Telephone, TX."
"A social media report indicated 3-inch diameter hail in Rockwall, TX."
"Emergency management reported golf-ball to baseball sized hail along FM 779 near the county line."
"A trained spotter reported Ping Ping ball sized hail one mile north-northeast of Plano, TX."
"A report via amateur radio indicated golf-ball sized hail in Krum, TX."
"A report received via amateur radio indicated golf-ball sized hail in the northeast part of Denton, TX."
"A trained spotter reported 2-inch diameter hail in Northeast Frisco, TX."
"Nickel to quarter size hail fell 4 miles east-northeast of Dalhart."
"A thunderstorm produced quarter size hail near I-10 and I-410 in Balcones Heights."
"Off-duty NWS employee reported quarter to golf ball sized hail along with damaging winds."
"Off-duty police chief reported a cracked windshield from tennis ball sized hail along with damaging winds, heav
"A fast moving hail storm caused multiple cars to slide off Highway 20 south in Rigby causing traffic to be bloc
"Media reported 3/4 inch hail near I-40 in western McDowell County."
"Storm chasers reported hail up to 3 inches in diameter fell 8 miles south-southeast of Masterson along Highway
"This hail was reported near I-35 and I-29 near Levee Road."
"This report was gathered via social media. The picture on Twitter showed a hail stone nearly one-half the size
"Two inch diameter hail occurred 8 miles south-southeast of Masterson."
"Tennis ball size hail fell 6 miles south-southeast of Masterson."
"Quarter size hail was measured at the National Weather Service office in Amarillo, located near Rick Husband Am
"Half dollar size hail was reported 5 miles north-northeast of Adrian. Accumulation of hail was reported."
"Most of the hail was dime-sized, but there were some hail stones up to the size of quarters. The wind gusts wer
"Wind-driven golf ball sized hail was covering the ground at TCU."
"A trained spotter reported penny-sized hail on Casper Mountain."
"Golf ball sized hail was reported at I-30 and Hulen Street."
"Emergency Management reported one-inch diameter hail in Hillsboro, TX."
"A thunderstorm rapidly intensified along and downstream of the Bighorn Mountain foothills northwest of Buffalo.
"Hen Egg sized hail was reported 6 miles south-southeast of Van."
"Amateur radio reported golf-ball sized hail near the intersection of Parker Road and Dallas North Tollway. The
"Golf ball sized hail reported 2 miles north-northeast of Plano. The hail shattered windows in cars, homes, and
"Golf ball size hail fell 6 miles east-northeast of Channing."
"Quarter size hail was reported 25 miles south-southwest of Perryton."
"Golf ball size hail was reported 9 miles south-southeast of Spearman."
"Golf ball size hail was reported 5 miles east-northeast of Channing."
"Trained Spotter reported quarter-sized hail along US 52 near Camp Coker Road."
"Lots of hail from pea size up to golf ball size was reported on Highway 385, 6 miles south-southeast of Hartley
"Quarter sized hail was reported 6 miles south-southeast of Van."

"Quarter size hail fell 2 miles east-southeast of Clarendon. This was the first of two storms that brought large
"Medstar reported around 2 inch hail damaged approximately 30 staff vehicles, 8 support vehicles, 20-30 ambulanc
"Golf ball size hail fell 5 miles west-southwest of Panhandle, along Highway 60 near County Road J."
"Half dollar size hail fell 8 miles north-northeast of Claude."
"A trained spotter reported golf ball sized hail approximately 5 miles south-southwest of Evant, TX."
"Off-duty NWS Employee reported ping pong ball sized hail."
"Very large hail up to 4 inches in diameter fell across portions of St. Charles County. The hardest hit areas s
"Quarter size hail fell 4 miles west-northwest of Goodnight."
"Larger than quarter sized hail was reported at residence north of MT-35 bridge over Flathead River. The hail pu
"The public reported quarter-sized hail from a thunderstorm west of Kalispell, MT."
"Quarter size hail was reported 25 miles south-southwest of Perryton."
"Hail up to two inches in diameter fell just southwest of Yarbrough, or about 5 miles east-southeast of Sturgis.
"Golf ball-sized hail was observed and photographed along Interstate 80 near Aragonite."
"Ping pong ball-sized hail was reported near Perry, Utah."
"Quarter to golf ball size hail fell from a nearly stationary storm in the Ashton area. The actual duration of h
"The public reported 1-inch hail via social media 3 miles south of Columbia Falls, MT."
"Off-duty NWS Employee reported half dollar sized hail."
"Off-duty NWS employee reported hen egg sized hail."
"Nickel to quarter size hail was reported 9 miles south-southwest of Dalhart."
"A social media report indicated quarter-sized hail near the intersection of Timberland and Beach Street in Kell
"A public report indicated half-dollar sized hail near the intersection of Yaggi Drive and China Berry Drive in
"A photo on social media confirmed that nickel-sized hail fell in Vero Beach South."
"This hail occurred near I-470 and NE Bowlin Road near Lakewood. This report was gathered via social media."
"Hail up to 2 inches in diameter was reported along state Highway 171 north of Kerrick,or about 10 miles south-s
"Several members of the public reported and took photographs of quarter-sized hail in West Valley City and West
"Nickel-sized hail was reported in Midvale."
"Public reported quarter size hail just north of I-40 near exit 111."
"A social media report indicated ping pong to golf-ball sized hail in Lewisville, TX."
"Multiple reports of ping pong ball- to golf ball-sized hail were received in West Valley City."
"Ping-pong ball sized hail was reported at Maverick Junction."
"Hail up to 3 inches in diameter damaged windshields of vehicles in Marsh on Highway 87, 7 miles south-southeast
"Hail was mostly nickel size with a few up to quarters. The actual duration of hail was likely at least several
"Ping pong ball sized hail 5 miles west-northwest of Celeste."
"Penny to quarter size hail was reported 5 miles west-southwest of Alanreed. The hail covered Interstate 40."
"Delayed report of hail upwards of ping-pong balls. Some damage to vehicles and crops."
"Delayed report, hail up to 2 inches reported at the Rock Springs 4-H Center."
"Half-dollar sized hail was reported along US-6 north of Delta. The hail broke the sunroof of a car, and also de
"One inch diameter hail was reported on I-85 at mile marker 134."
"Spotters reported quarter size hail at I-77 and Highway 150 and at Rinehardt Rd and Highway 150."
"Hail up to 1 inch in diameter occurred for a time near along Kolb Road near Escalante Rd, 29th St, and Golf Lin
"A #tSpotter Twitter report indicated quarter size hail fell at the I-40 and Highway 96 exit. Numerous other rep
"A member of the media reported quarter size hail on I-75 in Miramar. There were also multiple reports of pea si
"Hail up to nickel size covered the ground near the travel plaza at I-15 Exit 75."
"A public safety officer reported quarter sized hail, 6 miles south-southwest of Port St. Lucie."
"Dime-size hail covered the ground in and near the Somersett community in northwest Reno."
"Reported by Md-mg-5."
"Public reported nickel to quarter size hail across the north side of Spartanburg, from Asheville Highway to the
"Reported by WCAV-ch 19 Charlottesville."
"The public reported quarter size hail 5 miles south-southwest of Zephyr on County Road 258."
"The hail stones ranged from ping-pong to golf ball in size."
"A storm chaser observed wind-driven hail as large as tennis balls. The hail fell immediately west of what woul
"Reported by VA-fx-40."
"Estimated wind speeds of 50-60 MPH also occurred with the dime size hail."
"Estimated thunderstorm wind gusts of 65-70 MPH also occurred with the hail."
"Golf ball-sized hail was reported just west of West Point."
"Hail stones up to ping-pong ball size covered the ground."
"The largest hail stones ranged from ping-pong to golf ball size. The hail continued to fall into the following
"A trained spotter reported quarter size hail, 6 miles west-northwest of Potosi."
"A trained spotter reported half dollar to hen egg size hail falling along Texas State Highway 6, four miles nor
"The hail ranged from half dollar to ping-pong ball size."
"A trained spotter reported half dollar size hail, 8 miles east-southeast of Oplin."
"The fire department reported half dollar size hail, 3 miles north-northeast of Winchell."
"The largest hail stones ranged from ping-pong ball to hen egg size."
"Golf ball-sized hail was reported north of Duchesne."

"Nickel- to quarter-sized hail was reported in multiple locations across Utah County, including Payson and Salem
"Nickel- to quarter-sized hail was reported in Mt. Carmel, and the hail accumulated to a significant depth."
"Penny-sized hail was reported near Virgin."
"Penny-sized hail was reported in South Salt Lake."
"Nickel-sized hail was reported near St. George."
"Quarter-sized hail was reported in Kamas."
"There was a report of nickel-sized hail on the west shore of Lake Gogebic."
"A thunderstorm produced 3.5 inch hail near N. Foster Rd. and I-10 in eastern San Antonio."
"Quarter-sized (1) hail reported near Loysville."
"Trained spotter reported dime-to-nickel size hail covering the ground, along with 1.5 inches of rain within a 1
"Hail up to the size of nickels was reported in Thermopolis. Most of the hail was pea-sized and it accumulated u
"A spotter in Rockland observed penny-sized hail covering the ground."
"The Lemhi County dispatch reported ping pong ball-sized hail with the storm."
"NWS employee reported dime sized hail near the intersection of Miramar Parkway and Dykes Road from 545-550 PM E
"Estimated wind gusts of 30-40 MPH and almost an inch of rainfall accompanied the hail."
"Estimated wind gusts of 30-40 MPH accompanied the hail."
"A member of the public reported and took a picture of quarter-sized hail in Draper."
"Quarter size hail was reported 22 miles west-northwest of Mule Creek Junction."
"Half dollar size hail was reported a mile south-southwest of Whitaker."
"Penny size hail of 3/4-inch in diameter was reported just north of Cranford."
"Golf ball sized hail was reported about one mile east-southeast of the US Capitol."
"Slightly larger than quarter-size hail was reported."
"A severe thunderstorm produced golf-ball sized hail near Allenwood."
"A co-workers friend reports quarter size hail."
"Nickel size hail accompanied the straight-line winds. The hail lasted 15-20 minutes."
"A severe thunderstorm produced quarter-sized hail near Northern Cambria."
"A member of the public located 5 miles west-southwest of Lupfer reported ping pong size hail, which dented 3 ve
"A trained spotter took pictures showing 1 and 1/2 inch diameter hail one mile north-northeast of Walker."
"A caller reported nickel sized hail falling near I-75 and Bee Ridge Road in Sarasota."
"A sever thunderstorm produced nickel to quarter-sized hail in Lancaster."
"A severe thunderstorm produced quarter-sized hail in Airville."
"A severe thunderstorm produced quarter-sized hail in Loganville."
"A severe thunderstorm produced quarter-sized hail in southern Clinton County."
"The hail was nickel to quarter-sized."
"Broadcast media relayed a photo of hail estimated to be the size of at least golf balls near Spade. No damage
"The hail was ping pong to tennis ball-sized. No damage was reported."
"Time estimated--check on radar."
"A thunderstorm produced quarter size hail about a half mile north-northeast of Sabinal."
"Large hail and strong winds produced heavy damage to a residence eight miles east-southeast of Solano. Golf bal
"Dime to nickel size hail on I-40 west of Santa Rosa."
"Quarter sized hail occurred at KGNS-TV location in Laredo."
"A Crisp County Sheriff's Deputy reported quarter-sized hail along U.S. Highway 280 east of Cordele."
"A thunderstorm produced quarter size hail along I-35 near Jarrell."
"A Taliaferro County Sheriff's Deputy reported quarter-sized hail along I-20 near Crawfordville."
"Quarter size hail was reported 10 miles south-southeast of Sidney."
"Pea to quarter size hail covering the ground with heavy rain for 20 minutes south-southeast of Texline."
"Wind gusts to 40 mph accompanied the dime-size hail, which was reported from Bardwell to Cunningham."
"Hail covered the road but was sub-severe in maximum size. Minor street flooding was also reported."
"Golf-ball sized hail reported in downtown and the east side of Jackson. Approximately 20 to 30 patrol cars rece
"A coop observer traveling through the area reported nickel size hail covering the ground. However quarter size
"Hail to golfball-sized was reported in Empire, and up to three inches in diameter in Glen Arbor. Some vehicles
"Half-dollar size hail fell at the Olla Grocery near Highway 127."
"A resident reported quarter size hail that fell in the Vowells Mill community, via the KSLA-TV Facebook page."
"Quarter size hail was reported near Highway 109 just south of I-40."
"The hail ranged from quarter to ping-pong ball in size."
"Nickel size hail at I-40 and Eubank."
"Dime to nickel sized hail reported just south-southeast of Cloudcroft."
"Nickel-sized hail covered the ground in North Pekin."
"A swath of destructive hail up to golf ball size passed directly over O'Donnell accompanied by heavy rain and s
"Penny to quarter size hail along I-40 near mile marker 127."
"Wind-driven hail caused significant damage to the roofs and siding of 50 houses in Elmwood. One house suffered
"Reported at mile marker 393 on I-70."
"The hail ranged from quarter to ping-pong ball size and damaged vehicles, the house and garden."
"A picture of quarter size hail was posted to the KTRE-TV Facebook page."

```
"Quarter to Half-dollar size hail was reported in the Jordans Store community."
"A resident reported dime size hail falling in the Olla Community on the KNOE-TV Facebook page."
"Supercell thunderstorms (very rare for November in this area) affected much of El Paso. The strongest, which ac
"A resident reported dime size hail that fell in the Florien community, via the KSLA-TV  Facebook page."
"Trained spotter reported dime size hail and high winds on I-40 east near Gordonsville."
"There was a report just south of Pequaming of nickel-sized hail with one or two hailstones about one inch in di
"A spotter in Grand Marais reported dime-sized hail."
"There was a public report of dime-sized hail along West Bay in Grand Marais."
"A spotter just southeast of Grand Marais reported dime-sized hail."
"There was a public report of quarter-sized hail with some hail stones larger than quarter size.  The hail laste
"A spotter near Deer Park reported quarter-sized hail."
"The observer near Pelkie observed nickel-sized hail, and measured 0.75 inches of rainfall in less than a half a
"There was a public report of penny-sized hail four miles northwest of Manistique."
"There was public report of estimated one inch diameter hail west of Indian Lake.  The observer arrived onsite a
"There was a delayed public report of quarter-sized hail near Negaunee."
"Quarter-sized hail was observed at Stevens Field Airport."
"Ping pong ball size hail was reported five miles south-southeast of Federal."
"An off-duty NWS employee reported dime size hail near exit 2 of the Florida Turnpike."
"Quarter size hail was reported three miles south-southeast of Chadron."
"Half dollar size hail covered the ground seven miles west-southwest of Dalton."
"Quarter size hail was reported on Highway 20 about six miles east-southeast of Chadron."
"Golf ball size hail driven by straight-line winds estimated as high as 90 mph caused catastrophic damage to hor
"Quarter size hail covered the ground 12 miles west-northwest of Harrisburg."
"The observer near Jacobsville observed nickel-sized hail."
"There was a public report of penny-sized hail observed near Tapiola."
```

Perform Pre-processing

If you look at the descriptions above, you can see that some of them contain hyphenated words. Since you need to evaluate each word separately, start by replacing all hyphens with a space.

```
% Replace dashes with spaces so things like "quarter-sized" get
% counted as separate words
events.Event_Narrative = replace(events.Event_Narrative,"-"," ");
```

Now use the `tokenizedDocument` function to break each string into components.

```
documents = tokenizedDocument(events.Event_Narrative);
```

Visualizing the result with a word-cloud reveals lots of punctuation and stop words, which make it more difficult to identify useful features. Remove them using the `erasePunctuation` and `removeStopWords` functions.

```
% Erase punctuation.
documents = erasePunctuation(documents);
% Remove a list of stop words.
documents = removeStopWords(documents);
wordcloud(documents)
```

```
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
```

```
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
Warning: Graphics timeout occurred. To share details of this issue with MathWorks technical support, please
include that this is an unresponsive graphics client with your service request.
```


```
ans = 
  WordCloudChart with properties:

           WordData: ["hail"    "size"    "reported"    "sized"    "quarter"    "ball"    "fell"    "Quarter"    "ne
           SizeData: [3763 2132 1985 1015 703 641 625 609 530 438 436 343 300 288 279 276 262 252 246 246 241 216 21
    MaxDisplayWords: 100

  Show all properties
```

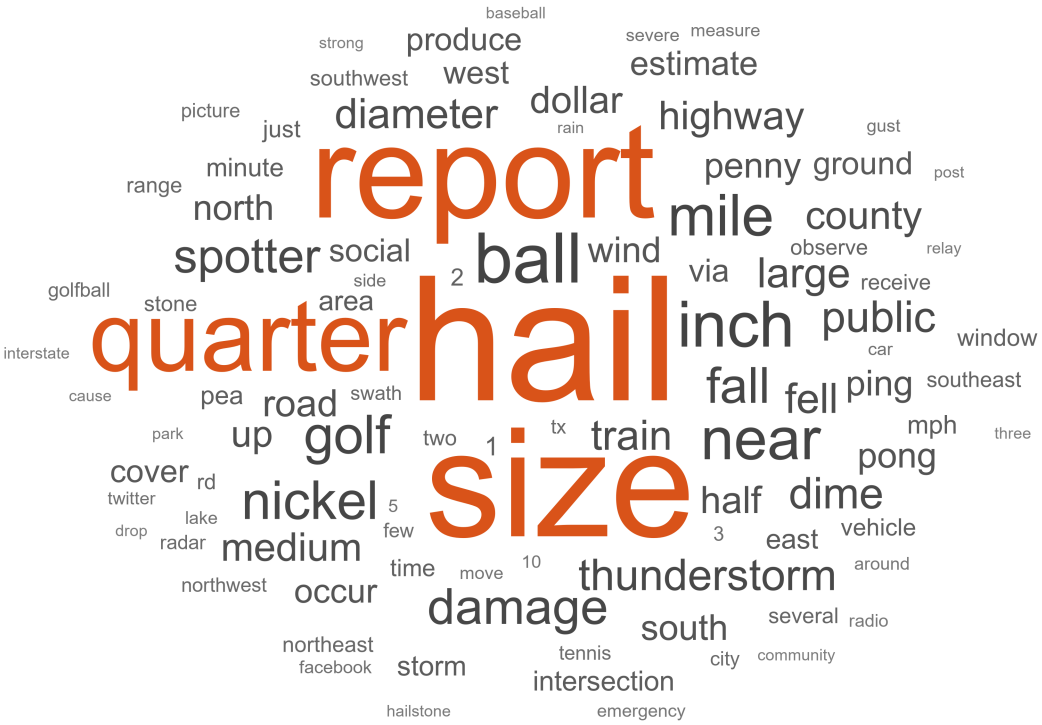
Find Features in Text Descriptions

Now take another look at the word-cloud above. You'll notice "size" and "sized" are both common words. These are good places to start looking for features since reference object words will often appear next to them. Use the `normalizeWords` function to reduce the words to their root form in lower case for easier searching.

```
documents = addPartOfSpeechDetails(documents);
documents = normalizeWords(documents,'Style','lemma');

% Update word-cloud
```

```
wordcloud(documents);
```



Next, you'll need to find word sequences containing "size".

Extract Multi-Word Sequences

Similar to the bag-of-words model you've seen previously, you can find two-word sequences or bigrams using the `bagOfNgrams` function.

```
bag2 = bagOfNgrams(documents,"NgramLengths",2)
```

```
bag2 =
  bagOfNgrams with properties:

          Counts: [4073×11935 double]
      Vocabulary: ["nickel"    "quarter"    "size"    "hail"    "fell"    "near"    "golfball"    "union"    "churcl
          Ngrams: [11935×2 string]
    NgramLengths: 2
       NumNgrams: 11935
    NumDocuments: 4073
```

```
bag2.Ngrams
```

```
ans = 11935×2 string
"nickel"     "quarter"
"quarter"    "size"
"size"       "hail"
"hail"       "fell"
"fell"       "near"
"near"       "alligator"
"golfball"   "size"
```

```
"fell"       "union"
"union"      "church"
"church"     "road"
     ⋮
```

You can then use this to extract all of the words that come immediately before all variations of “size”:

```
wordsBeforeSize = bag2.Ngrams(bag2.Ngrams(:,2) == "size",:)
```

```
wordsBeforeSize = 57×2 string
"quarter"    "size"
"golfball"   "size"
"measure"    "size"
"ball"       "size"
"dime"       "size"
"dollar"     "size"
"penny"      "size"
"up"         "size"
"baseball"   "size"
"nickel"     "size"
     ⋮
```

Notice the words for US coins like “quarter”, “dime”, “nickel”, and “penny”, sports items like “baseball” and “softball”, and food items like “pea” and “grapefruit”. You can also see words like “ball” and “dollar” and "pong". What words come before them? To see this, extract the three-word sequences or trigrams and look for the words preceding “ball size”, “dollar size”, and "pong size".

```
bag3 = bagOfNgrams(documents,"NgramLengths",3)
```

```
bag3 =
  bagOfNgrams with properties:

          Counts: [4073×16511 double]
      Vocabulary: ["nickel"    "quarter"    "size"    "hail"    "fell"    "golfball"    "union"    "near"    "220"
          Ngrams: [16511×3 string]
    NgramLengths: 3
       NumNgrams: 16511
    NumDocuments: 4073
```

```
wordsBeforeBallSize = bag3.Ngrams(bag3.Ngrams(:,2) == "ball" & ...
    bag3.Ngrams(:,3) == "size",:)
```

```
wordsBeforeBallSize = 5×3 string
"golf"       "ball"       "size"
"pong"       "ball"       "size"
"tennis"     "ball"       "size"
"billiard"   "ball"       "size"
"ping"       "ball"       "size"
```

```
wordsBeforeDollarSize = bag3.Ngrams(bag3.Ngrams(:,2) == "dollar" & ...
    bag3.Ngrams(:,3) == "size",:)
```

```
wordsBeforeDollarSize = 3×3 string
"half"       "dollar"     "size"
"report"     "dollar"     "size"
"hail"       "dollar"     "size"
```

```
wordsBeforePongSize = bag3.Ngrams(bag3.Ngrams(:,2) == "pong" & ...
    bag3.Ngrams(:,3) == "size",:)
```

```
wordsBeforePongSize = 1×3 string
"ping"       "pong"       "size"
```

There is “golf ball size", “tennis ball size", "billiard ball size", "half dollar size", and "ping pong size". You can probably safely assume that the word preceding "pong ball size" was "ping", but you can double check this as well.

Perform Custom Normalization

One thing you may have noticed above, is that both "nickle" and "nickel" showed up as alternative spellings for the same object. You can perform a custom normalization to deal with this situation by using the `replaceWords` function.

```
documents = replaceWords(documents,["nickle" ; "nickles"],"nickel");
```

Words like "golfball," "tennisball", "halfdollar", and "pingpongball" are not single words, but since you are more interested in finding the Ngrams than grammatical correctness, it may make it easier to substitute in the technically incorrect version and search for single "words":

```
% You can replace each rows of an input string array using replaceNgrams.
% In order to replace both "golf ball" and "golfballs" by "golfball", Use a
% 2x2 string array with an empty string in the 2nd row and column.
documents = replaceNgrams(documents,["golf" "ball" ; "golfballs" ""],"golfball" );
documents = replaceNgrams(documents,["tennis" "ball" ; "tennisballs" ""],"tennisball" );
documents = replaceNgrams(documents,["half" "dollar" ; "halfdollars" ""], "halfdollar" );
documents = replaceNgrams(documents,["ping" "pong" "ball" ; "ping" "pong" "" ; ...
    "pingpong" "ball" "" ; "pingpong" "" ""], "pingpongball" );
```

Find and Rank Search Terms

So far you've found and normalized a number of search terms corresponding to items in hail descriptions:

- US coins “quarter”, “dime”, “nickel”, “penny”, and "halfdollar"
- Sports items like “baseball”, “softball”, "golfball", "tennisball", "pingpongball", "billiard"
- Food items like “pea” and “grapefruit”

To order these items by size, you can find the object-to-size conversions used by the United States National Weather Service for each by searching online:

- https://www.weather.gov/media/pbz/skywarn/Hail_Chart.pdf
- https://www.nssl.noaa.gov/education/svrwx101/hail/
- https://www.spc.noaa.gov/misc/tables/hailsize.htm

Define a set of search words arranged by size

```
searchWords = ["pea","dime","penny","nickel","quarter","halfdollar","pingpongball",...
```

```
    "golfball","billiard","tennisball","baseball" ,"softball", "grapefruit"];
```

Now create a bag-of-words model to search for these terms.

```
bag = bagOfWords(documents);
```

Notice that you need to look for single "words" with a bag of words model since you normalized all multiple-word item descriptions to be single search terms. You could alternatively use a more complicated strategy by searching the bag-of-Ngrams models in addition to the bag-of-words model.

To find the indexes of these words in the `Vocabulary` variable of the bag-of-words model, you can use the `intersect()` function.

```
% Use intersect to find the indices of the searchWord in the vocabulary
% The 2nd output is the indices of the search terms in searchWords
% The 3rd output is the indices of the same terms in bag.Vocabulary
% The 'stable' indicates that the resulting orderings are in the same order
% as searchWords
[~,searchOrder,vocabOrder] = intersect(searchWords,bag.Vocabulary,'stable');
```

Now extract the equivalent "counts matrix" for just the search words:

```
searchWordCounts = bag.Counts(:,vocabOrder);
% some entries in searchWordCounts may be > 1
% get a matrix of logical values for where this is nonzero
hasSearchWord = full(searchWordCounts)>0 ;
```

Each row in this `hasSearchWord` corresponds to a different event, and each column to a search word. If the value at row i and column j is 1, then you know event i has search word j. You can get all of the ***(i,j)*** pairs using the `find` function.

```
[docIdx,wordIdx] = find(hasSearchWord);
```

Validate the search

Double check that a particular document / narrative has the search word your analysis predicts.

```
% Each docIdx and wordIdx are the position indices for all 1 entries in
% hasSearchWord, so there will be nnz(hasSearchWord) entries in each.
idx = 746; % choose any index from 1 to nnz(hasSearchWord)
searchWords(wordIdx(idx))
```

```
ans =
"nickel"
```

```
documents(docIdx(idx))
```

```
ans =
  tokenizedDocument:

   6 tokens: nickel dime size hail occur mamou
```

```
events.Event_Narrative(docIdx(idx))
```

```
ans =
"Nickel and dime size hail occurred in in Mamou."
```

Note that more than one search term may be in a document / event narrative. In the analysis below, such events will be double counted, e.g. if the narrative has both "quarter" and "golf ball". As an alternative, you could remove these rows, go with the larger / largest item, or some other strategy of your own. Try a couple of approaches and see if there is a significant difference in the results that follow. How many narratives does this apply to?

Compare property cost by hail size

Assemble a table with the property costs and sizes using the rows and columns of 1s in the `hasSearchWord` matrix you found previously.

```
% Focus on the events with nonzero cost
isNonZero =  events.Cost(docIdx) ~= 0;
% Create a hail table with the costs and labels
hail = events(docIdx(isNonZero),"Cost");
% Get categorical variable for sizes
hail.Size = categorical(searchWords(wordIdx(isNonZero)))';
hail.Size = reordercats(hail.Size,searchWords);
% Look at the assembled table
hail
```

```
hail = 396×2 table
```

	Cost	Size
1	2000	pea
2	100	pea
3	100	pea
4	10	pea
5	200	pea
6	2000	dime
7	5000	dime
8	1000	dime
9	500	dime
10	250	dime
11	200000	dime
12	10	dime
13	20	dime
14	200	dime
15	10	dime

	Cost	Size
16	5000	penny
17	1000	penny
18	50000	nickel
19	1000	nickel
20	500	nickel
21	250	nickel
22	500	nickel
23	100000	nickel
24	6000	quarter
25	5000	quarter
26	3000	quarter
27	5000	quarter
28	1000	quarter
29	5000	quarter
30	50000	quarter
31	10000	quarter
32	10000	quarter
33	10000	quarter
34	10000	quarter
35	2000	quarter
36	750	quarter
37	1000	quarter
38	1000	quarter
39	5000	quarter
40	50000	quarter
41	10000	quarter
42	30000	quarter
43	50000	quarter
44	1000	quarter
45	3000	quarter
46	1000	quarter
47	60000	quarter
48	2000	quarter

	Cost	Size
49	50000	quarter
50	3000	quarter
51	3000	quarter
52	12000	quarter
53	50000	quarter
54	1000	quarter
55	30000	quarter
56	3000	quarter
57	80000	quarter
58	1000	quarter
59	275000	quarter
60	275000	quarter
61	500	quarter
62	1000000	quarter
63	1000000	quarter
64	300000	quarter
65	3000	quarter
66	35000	quarter
67	1000	quarter
68	250000	quarter
69	25000	quarter
70	25000	quarter
71	6000	quarter
72	2000	quarter
73	500	quarter
74	1500	quarter
75	500	quarter
76	500	quarter
77	15000	quarter
78	20000	quarter
79	500	quarter
80	500	quarter
81	500	quarter

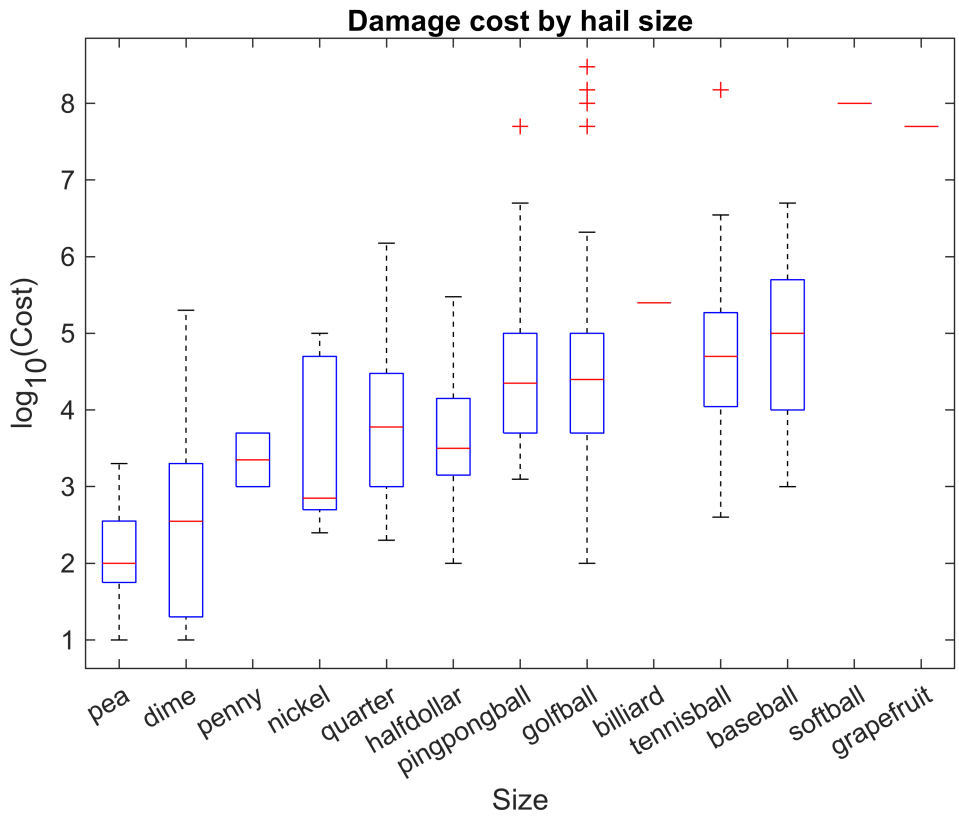
	Cost	Size
82	20000	quarter
83	20000	quarter
84	1000	quarter
85	10000	quarter
86	2000	quarter
87	20000	quarter
88	10000	quarter
89	10000	quarter
90	1150000	quarter
91	100000	quarter
92	10000	quarter
93	4000	quarter
94	10000	quarter
95	5000	quarter
96	5000	quarter
97	275000	quarter
98	110000	quarter
99	5000	quarter
100	200	quarter

⋮

Evaluate the Size Features Visually

If you take a look through the table, you can see the damage cost values vary over several orders of magnitude, so use a log 10 scale to visualize the results with the boxplot function.

```
boxplot(log10(hail.Cost),hail.Size)
ax = gca;
ax.XTickLabelRotation = 30; % rotate the tick labels so they don't overlap
ylabel("log_{10}(Cost)")
xlabel("Size")
title("Damage cost by hail size")
```



The chosen features do have some predictive value since there is a clear, overall trend of increasing damage with size. However, you can also see that some of these groups seem to have only one or two data points, e.g. "grapefruit", and some don't seem to be very different from each other, e.g. "penny" through "halfdollar". Re-grouping the sizes should yield a clearer trend.

Re-group the Sizes

Group "halfdollar" and smaller as "Small", "golfball" and "pingpongball" as "Medium", "billiard" through "baseball" as "Large", and since they're so much larger and have so few points, just remove "softball" and "grapefruit".

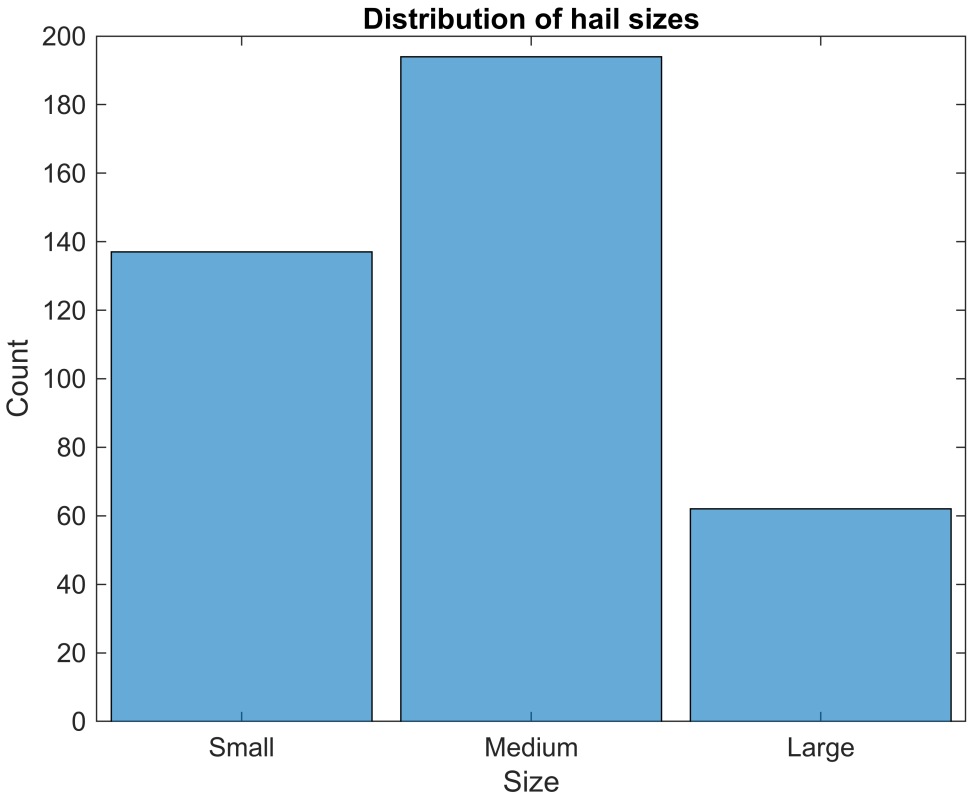
The following code was auto-generated using the live table editor to merge, rename, and remove categories. If you make changes to the search terms above, you'll have to edit or replace the code here.

```
hail.Size = mergecats(hail.Size,{'halfdollar','pea','dime','penny','nickel','quarter'});
hail.Size = renamecats(hail.Size,'halfdollar','Small');
hail.Size = mergecats(hail.Size,{'golfball','pingpongball'});
hail.Size = renamecats(hail.Size,'golfball','Medium');
hail.Size = mergecats(hail.Size,{'baseball','billiard','tennisball'});
hail.Size = renamecats(hail.Size,'baseball','Large');
hail.Size = removecats(hail.Size,{'softball','grapefruit'});
hail = rmmissing(hail);
```

Now you can take a look at the distribution of sizes.

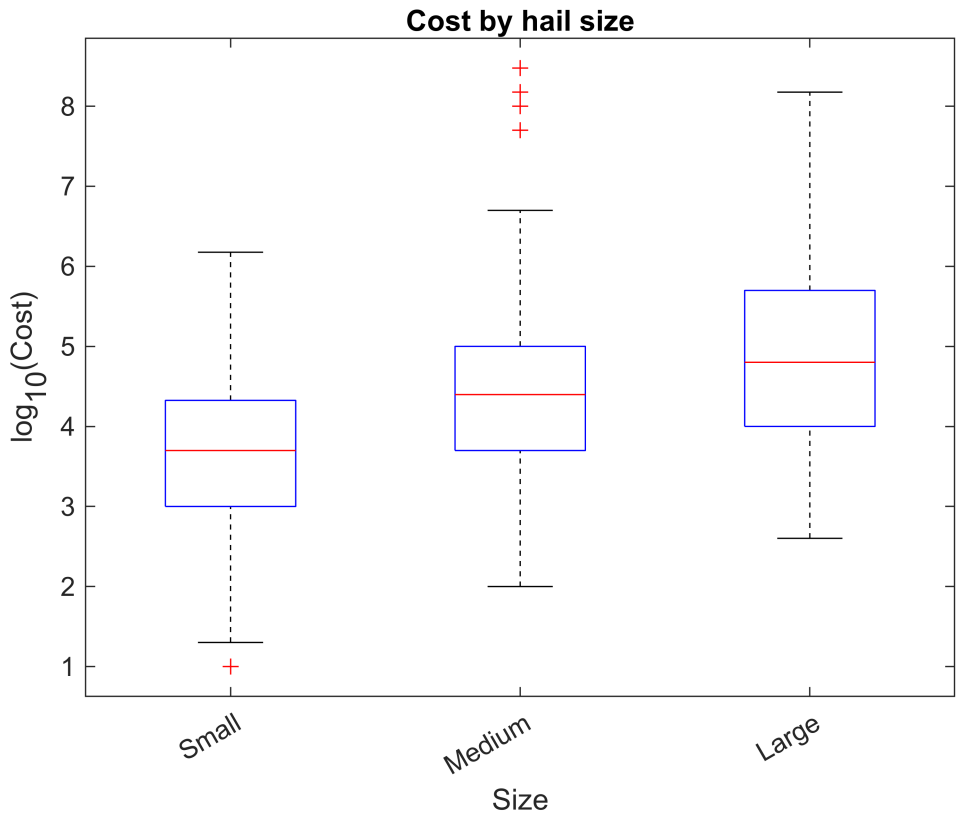
```
histogram(hail.Size);
xlabel("Size");
ylabel("Count");
```

```
title("Distribution of hail sizes");
```



Re-evaluate the Size Features

```
boxplot(log10(hail.Cost),hail.Size)
ax = gca;
ax.XTickLabelRotation = 30;
ylabel("log_{10}(Cost)")
xlabel("Size")
title("Cost by hail size")
```

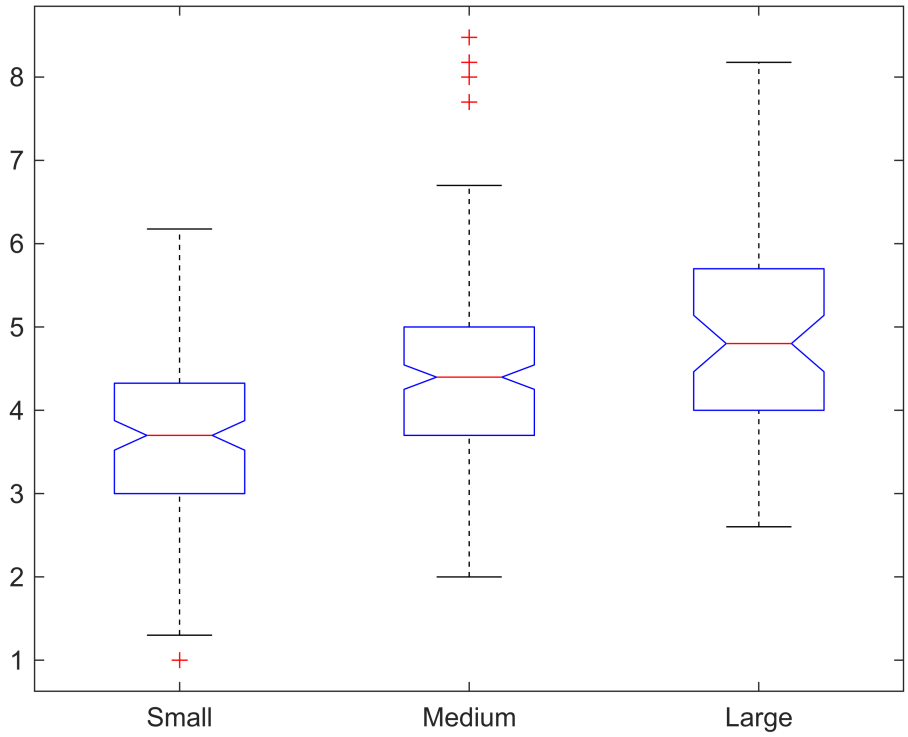


All the groups have a decent amount of data, and the trend looks a bit more clear now. However, are the groups really distinct enough to be good predictors of damage cost? It's time to evaluate the features beyond just visualization. To do this, you can use the `anova1` and `multcompare` functions to do a pair-wise ANOVA test with all the groups.

```
clf;
[p,t,stats] = anova1(log10(hail.Cost),hail.Size);
```

ANOVA Table

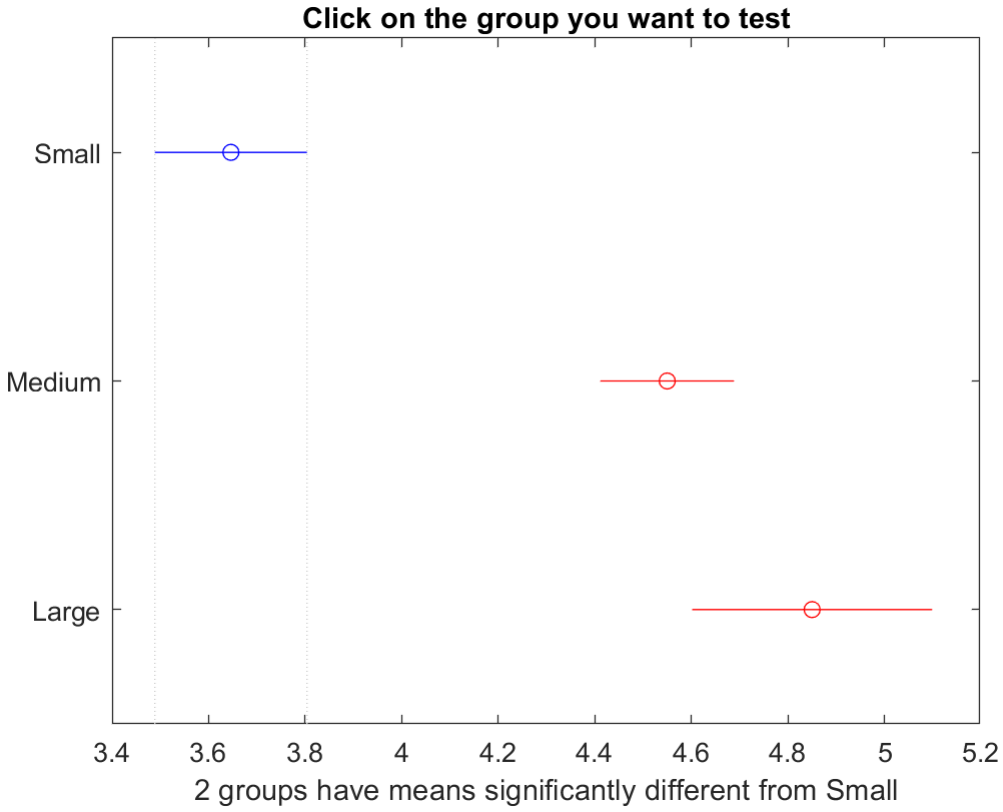
Source	SS	df	MS	F	Prob>F
Groups	89.369	2	44.6844	34.86	1.1766e-14
Error	499.882	390	1.2817		
Total	589.251	392			




```
figure;
multcompare(stats)
```

```
ans = 3×6
    1.0000    2.0000   -1.2001   -0.9040   -0.6079    0.0000
    1.0000    3.0000   -1.6104   -1.2042   -0.7981    0.0000
    2.0000    3.0000   -0.6873   -0.3002    0.0869    0.1637
```

```
% Generate the figure external to the live editor to interact with.
set(gcf,"Visible","on");
```



Unfortunately, while both "Medium" and "Large" are significantly different from "Small", they are not significantly different from each other. Wait a minute though! There were some significant outliers that might be biasing this analysis!

Clean Outliers from Size Features

To clean the outliers from each group, you can iterate through each category by name. Note that you concatenate the indices here because the table is sorted by category. Since the table was initially designed to be in order of smallest to largest size descriptors, merging categories maintained that ordering.

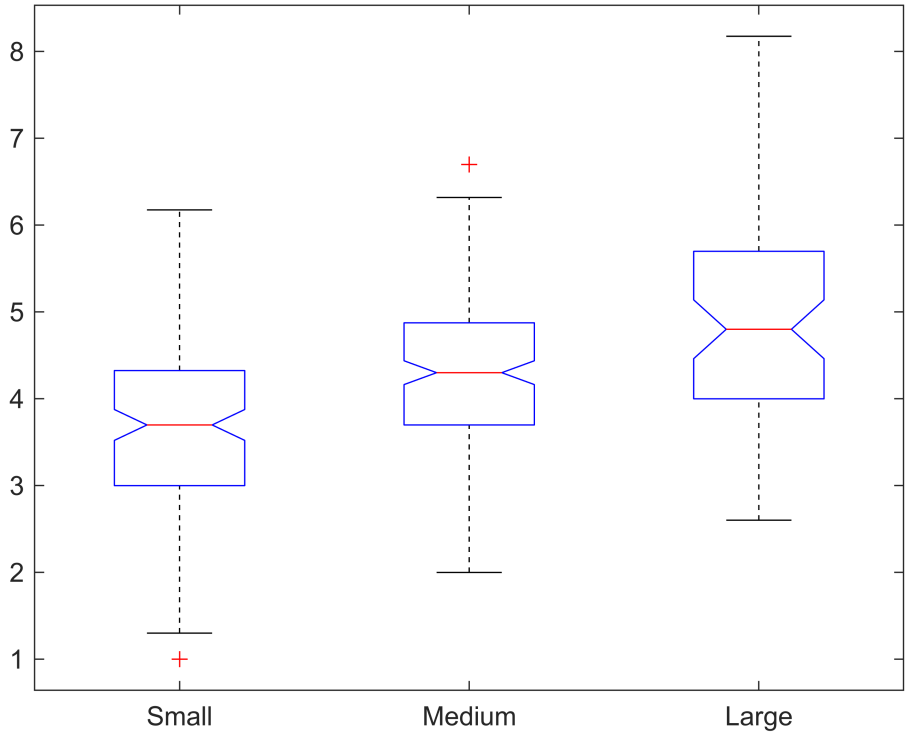
```
% Define an empty vector for the outlier indices
outlierIdx = [];
% Iterate through each category name (convert to string for comparison)
for catName = string(categories(hail.Size)')
    % Use logical comparison to extract a given size and find the outliers
    % and concatenate onto outlierIdx since the table 'hail' is sorted by
    % category
    outlierIdx = [ outlierIdx ; isoutlier( log10(hail.Cost( hail.Size == catName ))) ];
end
% Use the outlier indices to remove them from all groups
hail2 = hail(~outlierIdx,:);
```

Evaluate Cleaned Size Features

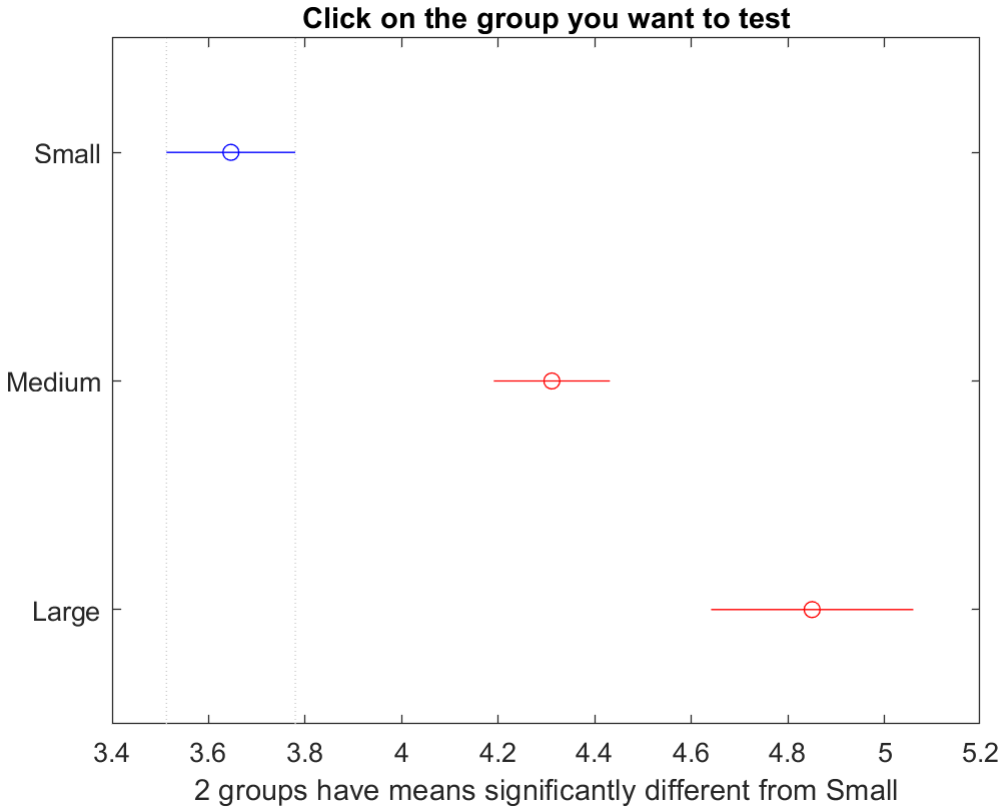
```
[p2,t2,stats2] = anova1(log10(hail2.Cost),hail2.Size);
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	69.857	2	34.9283	38.2	7.81476e-16
Error	344.695	377	0.9143		
Total	414.551	379			




```
figure;
multcompare(stats2);
% Generate the figure external to the live editor to interact with.
set(gcf,"Visible","on");
```



Now all three sizes are significantly distinct predictors of damage cost!

Summary

In this reading you've used the feature engineering workflow to process, explore, and extract then evaluate features for predictive cost modeling in text storm event descriptions. You performed a number of preprocess steps on the data in order to locate potential predictive features. Then you evaluated the features using both visual and statistical tools. You re-structured and cleaned the features further to improve their usefulness in a predictive model of storm damage cost.

Of course, you may have some ideas at this point on how to tweak, expand upon, or improve what you've seen here. A few additional suggestions beyond those already given above are:

- Repeat the same analysis using only property, or only crop cost. Do you get similar results?
- Try including the "softball" and "grapefruit" categories as "ExtraLarge".
- Try searching for additional terms or even removing some of those used initially.
- Use the actual numerical measurements for each object. See hailSizes.csv.
- Try finding additional size information not provided using reference objects, e.g. "1.5 inch". What pre-processing step would you need to change to accommodate this measurement in particular?
- Try using other years, or multiple years worth of data. What do you expect to result from using more data?