Summary of my work: I started with the question “what are the characteristics of a good movie”? I defined “good movie” to be one that had high Rotten Tomatoes and high IMDB ratings. I looked at US Gross, International Gross, Production Budget, rating, and genre, and I realized that this definition of “good movies” looked very similar to movies in general—my subset wasn’t very unique. So, I changed my definition of “good movie”—to one that had won an Oscar for Best Picture, Best Actress, Best Actor, or Best Director from 1990 – 2010. Again, these movies seemed to be fairly representative of movies as a whole (wide range of production budgets, grosses, etc.), so I changed my question one more time: what are the differences between these good movies? I compared the four types of winners, and found that Best Actress and Best Actor winners tended to cost and gross less than the Best Director and Best Picture winners.

Initial question: what are the characteristics of a “good” movie?
I’ll start by defining “good” as a movie that got high IMDB and high Rotten Tomatoes ratings. Let’s see what all our data look like
A good first step to answering this kind of question is to make a scatter plot looking at a few key “characteristics”. I had a ton of data (obviously), so I decided to apply a filter to limit my movies to those released between 1/1/1990 and 1/1/2012. These are the movies I’m most familiar with, and there’s the most data in this range. I also filtered my movies to the four most common ratings--G, PG, PG-13, and R. To make my graph clearer, I color coded by rating, and drew trend lines.

A few things jumped out about this visualization. There seemed to be a fairly consistent relationship between IMDB rating and rotten tomatos rating. But, there were also some notable outliers: movies that fared very well in Rotten Tomatoes ratings but very poorly in IMDB ratings, and movies that fared very well with IMDB but very poorly with Rotten Tomatoes. I’m interested in a third group, the movies that received very good ratings in both areas (I’ll call these the “good” movies). So, I’ll filter out this area and dig through that data.

Let’s look at the “good” movies—the ones rated highly by Rotten Tomatoes and IMDB
I visualized this subset of the data using a scatterplot. Color encodes rating and size encodes US gross. There is a notably high proportion of G rated movies.
Now, I've isolated the data for the “good” movies, so it's time to study their characteristics.

**PAUSE: what is an IMDB rating anyway?**

But, before I dig in, I want to look more in depth about how an IMDB rating is generated. It's generated through votes, and we have data about the number of votes each movie got, so I want to look at the relationship between votes and rating. Here's the relationship for all movies, filtered again by year (limiting 1990 – 2012) and rating (G, PG, PG-13, and R):

There's a trend here: we see a lot of variation in IMDB ratings for movies that didn't get a lot of votes--this makes sense, as a small sample can have a lot of variance. But as we move to the right along the x axis (more votes), the average rating becomes much higher. There are fewer movies in this area too. So, a lot of movies are getting a relatively low number of ratings, and there's a lot of spread in the average ratings they get. Fewer movies are getting a large number of ratings, and their average rating is much higher.

**Uh oh, what if our “good movies” by IMDB are just the movies that happened to get a lot of votes?**
Here’s the relationship between IMDB rating and IMDB votes for our subset of “good movies”.

<table>
<thead>
<tr>
<th>IMDB rating by IMDB votes for &quot;good&quot; movies</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPAA Rating</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>PG</td>
</tr>
<tr>
<td>PG-13</td>
</tr>
<tr>
<td>R</td>
</tr>
</tbody>
</table>

The plot of average of IMDB Rating (good movies) for IMDB Votes. Color shows details about MPAA Rating. The marks are labeled by Title. The data is filtered on MPAA Rating (good movies) as an attribute, which keeps G, PG, PG-13 and R. The view is filtered on IMDB Votes, which keeps non-Null values only.

Here, we see that our highly rated movies got a wide range of number of votes. This implies (although correlation doesn’t equal causation!!) that simply getting a large number of votes is not why our “good” movies had high IMDB ratings. Since we don’t have similar data about number of voters for the Rotten Tomatoes ratings, we can’t do a similar analysis with those data. But, we have shown that ratings for the “good” movies are consistently high regardless of the number of IMDB votes they got.

But still, what makes a good movie?
Could it be MPAA rating? Is it possible that the best movies tend to have a certain rating?

<table>
<thead>
<tr>
<th>Number of &quot;good&quot; movies with each MPAA rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPAA Rating</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>PG</td>
</tr>
<tr>
<td>PG-13</td>
</tr>
<tr>
<td>R</td>
</tr>
</tbody>
</table>

This graph makes it clear that far more “good” movies have an R rating. Interesting!

Here’s an interesting sidebar—most of the “good” movies are R-rated, but of the “good” movies, the real breadwinners are rated G. This graph shows the average US gross and worldwide gross by genre:

Average US and Worldwide Gross ($) by Rating, of "good" movies

<table>
<thead>
<tr>
<th>Average US Gross</th>
<th>Average Worldwide Gross</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPAA Rating</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
</tr>
<tr>
<td>PG</td>
<td></td>
</tr>
<tr>
<td>PG-13</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
</tr>
</tbody>
</table>
Are good movies the ones that MAKE a lot of money?
(this is a hard one to tease out--people may think that a movie is better because it's a high grossing movie, or movies might make a lot of money because people think they're good. Obviously, with this data set, we cannot tease apart correlation and causation). The visualization below shows the relationship between US gross and IMDB rating.

From this visualization, we can see that “good movies” are spread out across a wide range of grosses. So, good movies don’t make a lot of money.

The visualization using Rotten Tomatoes rating data instead of IMDB data for rating looks very similar:
Are good movies the ones that COST a lot of money?

To answer this question, I looked at my subset of “good movies” and graphed IMDB rating by production budget. Rating is color coded. It is clear that highly-rated movies span the full range of production costs.

The visualization using Rotten Tomatoes rating data instead of IMDB data looks very similar and makes the same point:

Here, we see that the answer is no—“good” movies aren’t just the ones that cost a lot to produce. Across all the highly rated movies, there is a wide spread of production budgets.
Do good movies come out at a certain time of year?
This visualization addresses that question; data for our subset of “good movies” are grouped by quarter. Stacked bars are color coded by rating:

Yes, it seems apparent that most of the “good movies” came out in the second two quarters. This makes sense with the progression of the year and Oscar season--movies that come out toward the end of the year are usually the ones considered for Oscars. Thus, it would make sense that some of the best movies come out in the 3rd and 4th quarters.

Do good movies tend to have the same genre?
For this visualization, I grouped our “good” movies by major genre, then plotted the number of movies falling into each genre. Rating is color-coded.

Indeed, it appears that drama outnumbers the other categories for “good movies.” Comedies come in second. It appears from this visualization that comedies and dramas appear to contain more R movies than the other genres, and this fits back with our earlier finding that more “good movies” were rated R.

At this point in our adventure we’ve learned a few things about the movies rated highly in both IMDB and Rotten Tomatoes:
1. They have a wide range of domestic and worldwide grosses (good movies don’t necessarily make extreme amounts of money either domestically or internationally)
2. They have a wide range of production budgets (money doesn’t necessarily buy success!)
3. They tend to have an R rating, but not necessarily.
4. They tend to come out in the later two quarters of the year
Here's most of this info (again, using our same data set: the movies rated most highly by both Rotten Tomatoes and IMDB, between 1990 – 2012, rated G – R) in one visualization, using the principle of small multiples:

**Good movies by quarter**

BUT, this is boring/uninteresting--it doesn't tell us what's different about these “good movies.” In fact, there are lots of movies both good and bad are rated R, that are dramas, which come out in the last two quarters. **I need to look at a better way to define “good movies.”** I'll define “good movies” as ones that have won an Oscar for Best Picture, Best Actor, Best Actress, or Best Director from 1990 to 2010.

**First, I'll look at some general trends for these movies:**
Data, filtered out by the parameters listed above (I looked up my own information for Oscars), is graphed below by creative type. Rating is color coded.

Again, this isn’t exciting or surprising--generally Oscar winners are more serious films, and they tend to be rated R. So, I’ll ask another question: **How profitable are these movies? How much do they cost to make?**
Data are filtered based on the criteria above. Each point reflects a movie’s production budget ($) and US gross ($). Rating is color coded.
Still, doesn’t appear to differ too much from movies in general. So, I’ll narrow in my question using this same set of data: How does production budget, domestic gross, and international gross vary for different types (Best Actress, Best Actor, Best Director, Best Picture) of Oscar winning (“good”) films? For example, do studios spend more on films that win Best Picture? Or perhaps they spend more money promoting individual actors or actresses? Let’s look at that:

Average production budget ($), US Gross ($), and worldwide gross ($) for “good” movies that did and did not win the Best Actress Oscar.

Average production budget ($), US Gross ($), and worldwide gross ($) for “good” movies that did and did not win the Best Picture Oscar.

There are some very interesting trends with these graphs. When it came to Best Picture, the winners tended to be the movies that studios spent more on and that made more money domestically and internationally. Best Actress winners had
the opposite pattern—Best Actress winners came from movies that were less well-funded and made less money domestically and internationally. There are many explanations for this pattern:

1. Voters felt disinclined to give the Best Actress Oscar and the Best Picture Oscar to the same movie. If studios invested more in getting their pictures to win Best Picture than winning Best Actress, this would make sense.
2. The roles that tend to lead to Best Actress victories come from lower budget, more independent films.
3. Best Picture films tend to be better overall (thus garnering more investment from studios and making more at the box office), while Best Actresses can shine in "less good" films.

Since we can't look at causation with this data set, I'll look at correlation, reaching my final question:

**How do movies that win Best Actress, Best Actor, Best Director, and Best Actress vary by Production Budget, US Gross, Worldwide Gross, and average number of awards won (out of these four categories).**

This visualization answers that question for only average number of awards won:

![Average Total Awards by Oscar Category](image)

The following visualization answers the question for all variables involved. Size encodes the average number of awards per movie (for example, if a movie wins four awards, that means it won Best Actress, Best Actor, Best Director, and Best Picture). Color encodes the award category (award category is doubly encoded as it is the focal point of the visualization). Production Budget, US Gross, and Worldwide Gross are layered so as to utilize the principle of small multiples and make the data easier to interpret.

**Production Budget ($), US Gross ($), and Worldwide Gross ($) by Oscar Category**

![Production Budget, US Gross, Worldwide Gross by Oscar Category](image)
This graph shows a number of interesting final results:

1. Best Actress-winning movies tend to have the lowest worldwide gross, lowest US gross, and lowest production budget. Best-Actor winning movies are almost as low, but slightly higher. Best Director and Best Picture both have a larger spread of grosses and budgets.

2. Best Actress and Best Actor-winning movies tend to have fewer overall awards, while best director and best picture movies tend to have more.
(1) Let’s find out more about the movies that got good Rotten Tomatoes ratings but bad IMDB ratings.
To visualize this, I made a scatterplot that isolated the movies in question. Rating is color coded, and US gross is encoded by size.

There are a few things to notice here—first, there is a big spread of MPAA ratings (although the majority of movies have an R rating). There’s also a large spread of US gross—with some movies like Raise the Titanic making very little and others like Beauty and the Beast making a lot. Just from glancing at the titles, these seem like movies that might be very popular among the general public, but not very critically acclaimed.

(2) But what about the other outliers, those with high IMDB ratings and low Rotten Tomatoes ratings?
I visualized these this subset of the data in the same way I visualized the low IMDB/high Rotten Tomatoes data. Here’s the graph:
Movies with a high IMDB rating and a low Rotten Tomatoes rating.

Notably, there are no G rated movies in this data set.