Homework 7 中央委员数据可视化

徐佳怡　　516021910396

数据描述与分析:

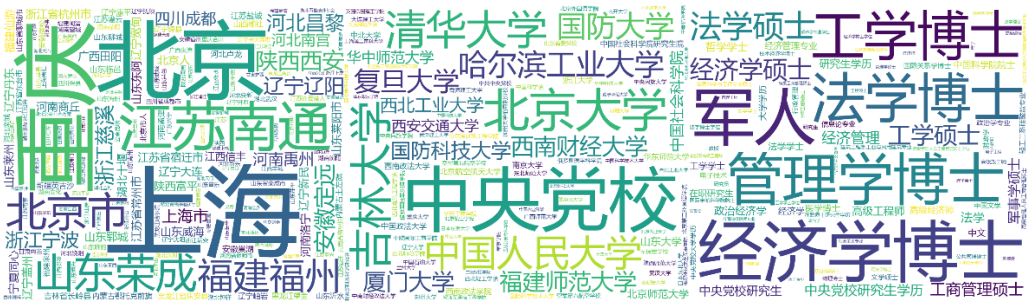
本次作业数据包含 204 位委员的性别，民族，出生地，出生日期，毕业院校，专业背景，担任职务，以及具体履历信息。其中，前 7 种信息可以归类为属性特征，最后一种履历信息描述了委员们的生涯轨迹。为了具体分析委员们的调任升迁和共事情况，我们可以将各个委员的履历进行切分，最终得到 3246 条履历信息。

我认为这个数据集可以从以下角度分析：

1. 性别比例：从数据我们能很轻易地看出委员绝大部多数（>95%）都是男性。所以，虽然性别比例是一个很重要的信息，但是在可视分析时需求不大。
2. 委员们的民族信息可以反映出人们的代表是否具有多样性。我用 python 的 wordcloud 库做了一个委员的民族词云图，如右图所示：

图中显示汉族委员占据大多数，但蒙古族，维吾尔族等族均有不少委员，这说明了中国是个包容各民族的国家。

3. 出生地、毕业院校，专业背景等信息都可以反映出哪些地区，哪些学校，哪些专业盛产中央委员。同样，我用 python 的 wordcloud 库对这三类信息分别做了词云，如下图所示：

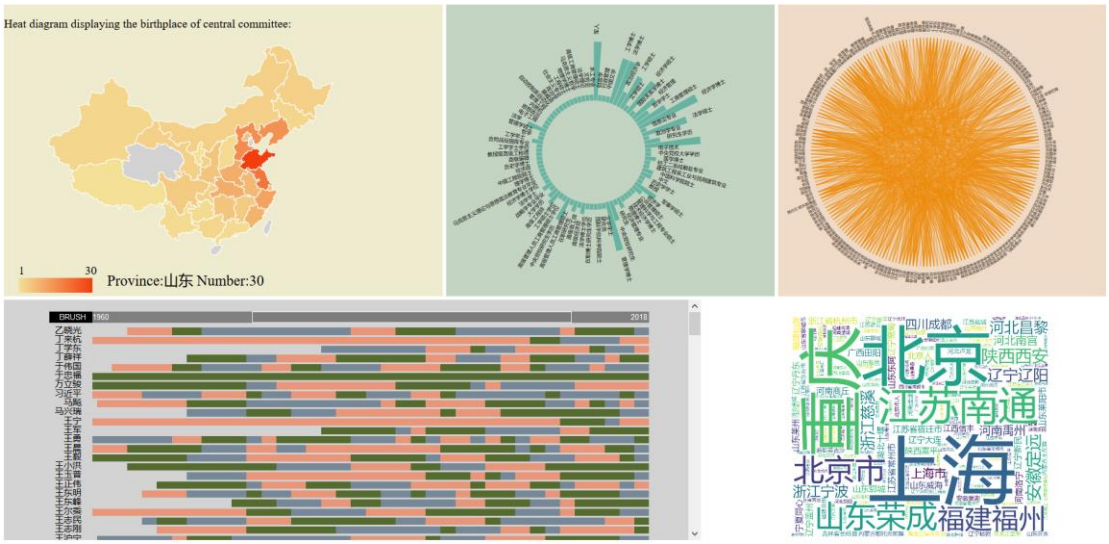


由图可见，北京，上海，山东等地是很多中央委员的出生地；中央党校，北京大学，清华大学，中国人民大学，吉林大学等学校培养了大批中央委员；绝大部分中央委员是军人，经济学博士，管理学博士，法学博士，工学博士。

4. 对中央委员的履历进行分析，从中可以获取委员的共事信息和生涯轨迹。

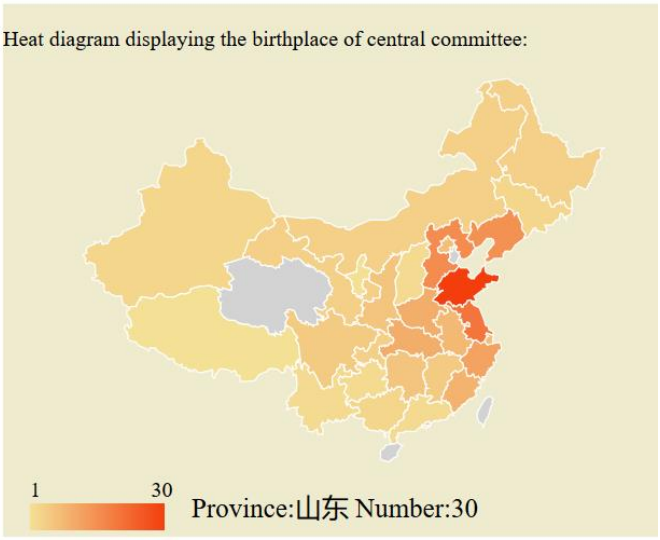
设计宗旨及过程：

下图是我的可视化作品总览：



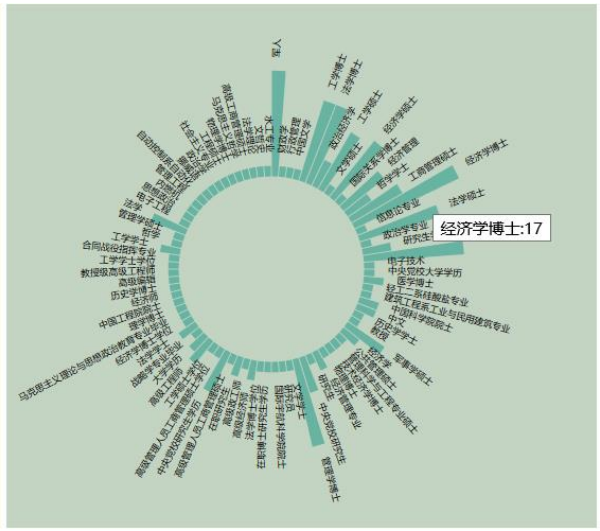
1. 首先，我想展现委员们的出生地情况，为了更清晰直观的展现，我选择了在地图上通过热力图展现各省份委员的人数。由于颜色并不能精确表示人数的数字信息，我在交互上设置了鼠标滑动就在地图下方显示省份及人数的文字信息.

   好处：这张地图能清晰直观地展现委员的出生地分布情况，也能通过交互精确的展示出生地分布数字信息。



   难点：由于地图上一次作业画过，所以这次画并没有太大的难度。主要难度在于数据的处理。一个问题是中文的乱码问题，需要通过设置 encoding 的格式解决。为了统计每个省份的委员人数，我们可以统计出生地中每个省份名字出现的频数，这直接通过字典（dict）统计就可获得。

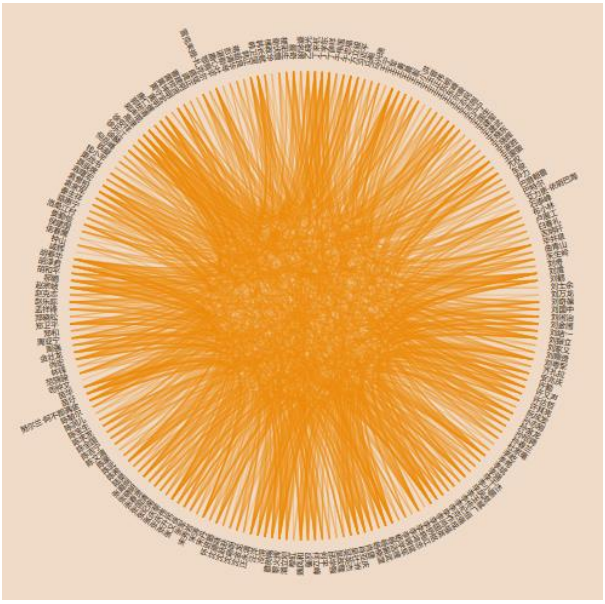
2. 其次，我想展现委员们的专业背景信息。我统计了每种专业对应的委员人数。这种信息通过普通柱状图就能展现。为了做一点创新，我选择了美观度更高一些的环形柱状图。将鼠标移至柱状图上，就能显示详细的专业背景人数。



好处：环形柱状图能很清晰地展现各种专业背景地委员人数，而且相较于普通柱状图更为美观。

难点：这张图地难点，主要在于环形柱状图地理解和绘画，例如如何设计离圆心越远，柱越粗地效果。数据处理过程要注意字段地切分，要将切割符号“,”“、”统一转成“，”，然后再运用 split()将一个人的多重专业背景分开，便于最终的统计。

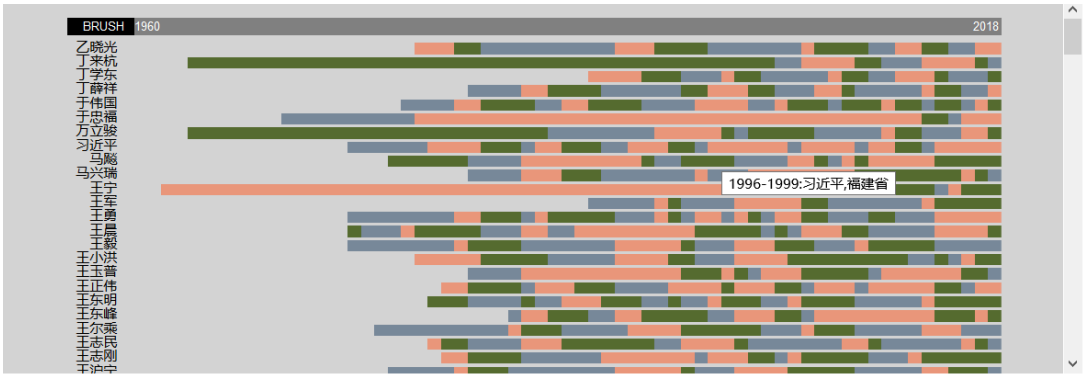
3. 再次，我想展现各个委员是否曾经共事的信息。这种信息需要链接图来展现。我们可以将委员们排在圆形上，然后用连线来表示两个委员之间是否存在共事经历。这种图叫做 edge bundling.



好处：能够直接通过连线表示委员们的共事信息，直观。

难点：数据处理较为困难。我根据已经处理好的委员履历信息，判断两个委员是否拥有两条对应的履历，其中开始结束时段有交集，并且工作单位相同，有的话则认为这两个委员曾经共事过，给他们添上连线。另一大难点就是画图，需要理解 edge bundling 具体的绘图逻辑。

4. 最后，我想展现各个委员具体的履历信息，所以我绘制了一张 timeline 图。每一个色块代表一位委员的一段经历。色块的起始和结束位置代表了经历的开始和结束时间。另外，我还为 timeline 图添加了一个 brush，用户可以刷选时间，展现某一段时间委员们的履历信息。将鼠标移至色块上，就能查看详细的时间，委员以及职位的详细信息。

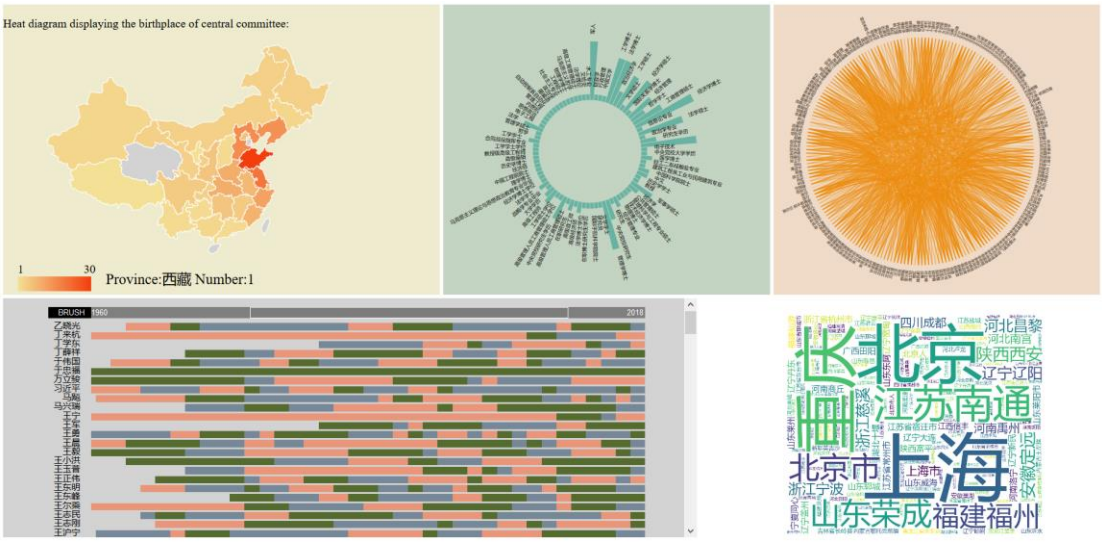


好处：从 timeline 图中，我们可以看出委员们的从业时间长短；是快速调任升迁还是长久的担任某一职务等信息。通过 brush，我们能限制履历的时间范围。

难点：timeline 图最大的难点在于 brush 的设置。通过 brush,我们可以得知限制的数据范围，通过这个范围信息，我们就能对原有的完整信息进行筛选，选出那些在所选范围内的数据进行展示。另外一个难点在于数据的展示。因为有 204 个委员，timeline 的长度很长，为了在有限的范围内展现，我们可以通过添加下拉框来限制。

可视化结果:

我的可视化作品最终如下:



从这个可视化中，我有如下发现:

1. 东部地区例如山东，江苏和浙江有大量的中央委员，这可能和这些地区的人口密度大，经济发展好，教育程度高，离北京较近有很大的关系。
2. 在中央委员中，专业背景是军人，经济学博士，工程博士，经济学博士，管理学博士，法学博士占比很大。从中可以看出，中央委员们普遍文化素质很高，大多都是博士学历。另外，法学，经济学，管理学和中央管理决策关系较大，所以中央委员专业背景是它们的很多。

3. 中央委员们普遍存在共事经历。我猜想成为中央委员也和曾经共事的伙伴有很大的关系，可能就是那些共事伙伴们的相互激励和影响，才让他们最终在中央委员的舞台上相见。
4. 从时间线图可以看出，大多数委员前几份履历的时间都比较长，这是经验积累的阶段，所以历时较久，在成为中央委员之前的几段经历都会较为短暂，我猜想可能是能力较高，在职位上表现优秀，就被迅速升职。由此，我们年轻人也要知道，我们要耐得住寂寞，在刚开始工作的时候，不能很浮躁，一心想着跳槽升职，要在一开始的阶段沉下心来，好好积累经验，之后机会自然会到来。

感受：

在这次作业中，我真实地感受到了什么是数据可视化，先有数据分析才有可视化，而且数据分析才是可视化地出发点和目的点。我们由初步的数据分析得出我们对现象的猜想，然后通过可视化检验我们的猜想是否正确，并将结果进行直观的展示。通过可视化的展现，我们可以进一步修正我们的猜想，并对数据进行更深刻的分析。可视化和数据分析是相辅相成，并且应该是循环往复的过程。

先前一次作业锻炼的是我们可视化画图的能力，这次作业就是锻炼了我们数据可视化的分析能力。我在这次作业中体会到了数据分析的不易。但是现实生活中的数据就是纷繁复杂的，就是要通过一步步的分析处理才能得到可用的数据。这次作业可谓是让我从理想环境向现实环境迈出了一大步。